MAIN INSTRUMENT SEPARATION FROM STEREOPHONIC AUDIO SIGNALS USING A SOURCE/FILTER MODEL

Jean-Louis DURRIEU[1], *Alexey OZEROV*[1], *Cédric FÉVOTTE*[2], *Gaël RICHARD*[1] *and Bertrand DAVID*[1]

[1]Institut Télécom; Télécom ParisTech; CNRS LTCI
37 rue Dareau, 75014 Paris, France

[2]CNRS LTCI; Télécom ParisTech,
37 rue Dareau, 75014 Paris, France

Email: firstname.lastname@telecom-paristech.fr - web: http://perso.telecom-paristech.fr/durrieu/en/eusipco09/

ABSTRACT

We propose a new approach to solo/accompaniment separation from stereophonic music recordings which extends a monophonic algorithm we recently proposed. The solo part is modelled using a source/filter model to which we added two contributions: an explicit smoothing strategy for the filter frequency responses and an unvoicing model to catch the stochastic parts of the solo voice. The accompaniment is modelled as a general instantaneous mixture of several components leading to a Nonnegative Matrix Factorization framework. The stereophonic signal is assumed to be the instantaneous mixture of the solo and accompaniment contributions. Both channels are then jointly used within a Maximum Likelihood framework to estimate all the parameters. Three rounds of parameter estimations are necessary to sequentially estimate the melody, the voiced part and at last the unvoiced part of the solo. Our tests show that there is a clear improvement from a monophonic reference system to the proposed stereophonic system, especially when including the unvoicing model. The smoothness of the filters does not provide the desired improvement in solo/accompaniment separation, but may be useful in future applications such as lyrics recognition. At last, our submissions to the Signal Separation Evaluation Campaign (SiSEC), for the "Professionally Produced Music Recordings" task, obtained very good results.

1. INTRODUCTION

Applications of musical source separation such as [1] or [2] show the relation between Music Information Retrieval (MIR) topics and source separation, especially in order to achieve instrument classification on polyphonic/poly-instrumental recordings or melody and drum transcription. While source separation may improve the results in classifying instruments or transcribing melody or drum tracks, these MIR results can also help in the separation process. There are numerous examples of the use of this interrelation between these 2 fields, e.g. [3] and [4] mainly focus on solo/accompaniment extraction, and [5] uses solo enhancement to improve lyrics recognition. However, most of the studies in de-soloing applications, to our best knowledge, deal with mono-channel signals, while most of the music recordings, nowadays, are at least stereophonic. On the other hand, multi-channel music source separation is a widely studied subject, e.g. in [1] or [6], but the existing methods hardly generalize to vocal signals.

We propose in this article to extend the monophonic model proposed in [4] to the case of stereophonic signals, i.e. an audio solo/accompaniment separation algorithm adapted for signals with a solo voice played by a harmonic instrument, e.g. a singer or any wind instrument. The separation principle follows an iterative method: we first roughly estimate the parameters corresponding to the main instrument and to the accompaniment, then we detect a main melody stream and estimate during a second round the parameters so that they better fit the chosen melodic line. The separation itself is held thanks to Wiener filters as in [7]. We also propose two extensions to the original solo source/filter model from [4]: an explicit model of the smoothness of the filter part and a model for the unvoiced parts of the solo instrument, which requires a third round of parameter estimation. The proposed model allows to estimate the parameters jointly from the left and right channels of the signal, and thus takes advantage of inter-channel intensity differences (IID). Two systems based on the proposed model were evaluated at the SiSEC campaign, for the "Professionally Produced Music Recordings" separation task [8], and obtained very good and promising results.

This article is organized as follows: in section 2, we introduce the stereophonic signal model which extends our previous works on solo/accompaniment separation. The method, i.e. the Maximum Likelihood (ML) criterion and the resulting updating rules, is presented in section 3. At last, results on the new features of the system, i.e. filter smoothness, unvoicing and stereophony, are reported and discussed in section 4. We conclude with future perspectives in section 5.

2. STEREOPHONIC SIGNAL MODEL

2.1 Modelling the stereophonic observation signal

Let us consider an observed stereophonic sampled audio signal $[x_L(t), x_R(t)]^T$, where t is the sample number, x_L (resp. x_R) is the signal from the left (resp. right) channel. The $F \times N$ Short-Time Fourier Transforms (STFT) of both channels are respectively denoted X_R and X_L, with F the number of frequency bins and N the number of analysis frames. The STFTs are assumed to be the instantaneous mixtures of two contributions, the solo part V (for "Voice") and the accompaniment part M (for "Music").

The stereophonic aspect is modelled as a simple "panning" effect: the original sources are assumed monophonic and mixed together into the stereophonic signal by applying different amplitude levels for each channel to simulate their spatial positions. The solo V is further assumed to have only one static spatial position and the accompaniment to be modelled with J several components, each of which have their own static spatial position. We assume that the STFTs, at frequency f and frame n, are given by:

$$\left\{ \begin{array}{rcl} X_{R,fn} & = & \alpha_R V_{R,fn} + \sum_{j=1}^{J} \beta_{Rj} M_{Rj,fn} \\ X_{L,fn} & = & \alpha_L V_{L,fn} + \sum_{j=1}^{J} \beta_{Lj} M_{Lj,fn} \end{array} \right. \quad (1)$$

where V_R, V_L, M_{Rj} and M_{Lj} are supposed to be realizations of random variables (r.v.). We assume that these r.v. are all mutually independent and individually independent across both frequency and time.

We exploit the stereophonic information only by considering that the signals for both channels (left and right) for one contribution V

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and partly funded by the French ANR Project SARAH (StAndardisation du Remastering Audio Haute-definition).

or M share the same statistical characteristics:

$$\left.\begin{matrix} V_{R,fn} \\ V_{L,fn} \end{matrix}\right\} \sim \mathcal{N}_c(0, S_{V,fn}) \text{ and } \left.\begin{matrix} M_{Rj,fn} \\ M_{Lj,fn} \end{matrix}\right\} \sim \mathcal{N}_c(0, S_{M,j,fn}) \tag{2}$$

where $\mathcal{N}_c(0,\sigma_Z^2)$ stands for complex proper Gaussian distribution with probability density function (pdf):

$$N_c(Z;0,\sigma_Z^2) = \frac{|Z|}{\pi\sigma_Z^2}\exp\left(-\frac{|Z|^2}{\sigma_Z^2}\right). \tag{3}$$

$S_{V,fn}$ and $S_{M,j,fn}$ are the variances for the solo signal and for the j^{th} component of the accompaniment, at frequency f and frame n. These variances are explicitly parameterized as in [4] as explained in section 2.2.
The resulting stereophonic signal therefore is distributed as follows:

$$\left\{\begin{array}{rcl} X_{R,fn} & \sim & \mathcal{N}_c\left(0, \alpha_R^2 S_{V,fn} + \sum_{j=1}^J \beta_{Rj}^2 S_{M,j,fn}\right) \\ X_{L,fn} & \sim & \mathcal{N}_c\left(0, \alpha_L^2 S_{V,fn} + \sum_{j=1}^J \beta_{Lj}^2 S_{M,j,fn}\right) \end{array}\right. \tag{4}$$

Our model is closely related to classical multi-channel source separation for instantaneous mixtures. Let $X_{fn} = [X_{R,fn}, X_{L,fn}]^T$ be the stereophonic observation vector and

$$A = \left[\begin{array}{ccccccc} \alpha_R & \beta_{R1} & \dots & \beta_{Rj} & \dots & \beta_{RJ} \\ \alpha_L & \beta_{L1} & \dots & \beta_{Lj} & \dots & \beta_{LJ} \end{array}\right]$$

the mixing matrix. If we drop the indices R and L in the right-hand side of equation (1), the model can be re-written as:

$$X_{fn} = \left[\begin{array}{cccc} \alpha_R & \beta_{R1} & \dots & \beta_{RJ} \\ \alpha_L & \beta_{L1} & \dots & \beta_{LJ} \end{array}\right] \left[\begin{array}{c} V_{fn} \\ M_{1,fn} \\ \dots \\ M_{J,fn} \end{array}\right] = A \left[\begin{array}{c} V_{fn} \\ M_{1,fn} \\ \dots \\ M_{J,fn} \end{array}\right] \tag{5}$$

The observed signal is the instantaneous mixture of all the contributions V_{fn}, $M_{j,fn}$ for $j \in [1,J]$. Such a model was used in [9]. We however chose the model described by equation (1) because it leads to a simpler parameter estimation. Such a model may also be more robust to modelling errors, especially when the observed signal is not an instantaneous mixture.
Another difference between the proposed system and [9] is also the specific model we use for the solo part. The technique in [9] is blind and principally uses the spatial directions to cluster the basis spectra into different sources. The proposed system uses a production model for the main instrument in order to identify and separate it. In audio recordings, several instruments can indeed be mixed with the same direction: both the singer and bass tracks could, for instance, be panned "in the middle". In such a case, the proposed algorithm will be able to separate these two sources, while [9] would need human intervention.

2.2 Solo part source/filter model

We assume that the solo part is played by a monophonic and harmonic instrument, e.g. a human singer. We use the source/filter model proposed in [4], which is well adapted to this type of signal and we integrate an additional smoothness constraint on the filter frequency responses inspired by [10] as well as a model for the unvoiced parts of the solo voice.
For the variance $S_{V,fn}$ of the solo voice, the speech-processing inspired source/filter model already obtained good results [4]. It allows the algorithm to seek for harmonic signals, thanks to a glottal source model on the source part, while still able to adapt the amplitudes of the different harmonics in the spectral comb through the filter shape estimation. The variance is parameterized as follows:

$$S_{V,fn} = S_{\Phi,fn} S_{F_0,fn}$$

with $S_{\Phi,fn}$ and $S_{F_0,fn}$ respectively the filter and the source contributions to the variance. We denote the $F \times N$ variance matrices S_V, S_Φ and S_{F_0} the matrices whose entries respectively are $S_{V,fn}$, $S_{\Phi,fn}$ and $S_{F_0,fn}$.
A unified voiced/unvoiced source model: the source variance is modelled as a non-negative linear combination of the spectral combs of all the N_{F_0} possible (allowed) fundamental frequencies. These spectra form a $F \times N_{F_0}$ matrix W_{F_0}. The associated amplitude coefficients form a $N_{F_0} \times N$ matrix H_{F_0} such that :

$$S_{F_0} = W_{F_0} H_{F_0} \tag{6}$$

It is worth noting that this formalism allows for a wide range of possibilities. It was first designed to fit voiced parts of the solo instrument [4] but one can add another "unvoiced" basis vector to W_{F_0} and set it to a uniform value for all the frequencies: it then models the expected source part for unvoiced sounds. As described in section 3.4, we prefer to estimate this unvoiced part in an additional round of parameter estimation, in order to avoid catching too many other "noisy" components, e.g. drums, which do not correspond to the main (melodic) instrument.
A smoothing strategy for the filter part: similarly, we define the $F \times K$ filter dictionary matrix W_Φ, where K is the number of different filter shapes that are allowed. The activation coefficient for the resulting filters in S_Φ form the $K \times N$ matrix H_Φ such that $S_\Phi = W_\Phi H_\Phi$. To model the smoothness of these filter frequency responses we introduce a $F \times P$ dictionary of smooth "atomic" elements W_Γ. Each filter, i.e. each column vector from W_Φ, is then decomposed on this basis, as being a non-negative linear combination of the column vectors of W_Γ. We then define the $P \times K$ matrix H_Γ such that: $W_\Phi = W_\Gamma H_\Gamma$. By construction, each filter in W_Φ is the sum of smooth functions, and is therefore also smooth.

$$S_\Phi = W_\Phi H_\Phi = W_\Gamma H_\Gamma H_\Phi \tag{7}$$

At last, the variance matrix S_V for the solo part is parameterized as follows, where the dot '.' represents element-wise product between the matrices:

$$S_V = S_\Phi . S_{F_0} = (W_\Gamma H_\Gamma H_\Phi).(W_{F_0} H_{F_0}) \tag{8}$$

We will refer to this solo voice model (and by extension to the complete solo/accompaniment model) as the "Instantaneous Mixture Model" (IMM), in contrast with the "Gaussian Mixture Model" (GMM) as used in [11]. Indeed, the source part can be seen, in its temporal counter-part, as the instantaneous mixture of all the possible notes, with amplitudes corresponding to the activation coefficients H_{F_0}.

2.3 Accompaniment model

As in [4], each of the J components for the accompaniment music is modelled as a centered Gaussian whose variance at frequency f and frame n is:

$$S_{M,j,fn} = w_{fj} h_{jn}$$

For each channel $C \in \{R,L\}$, we can therefore compute the global variance for the accompaniment $S_{M_C,fn}$:

$$S_{M_C,fn} = \sum_{j=1}^J \beta_{Cj}^2 S_{M,j,fn} = \sum_{j=1}^J w_{fj} \beta_{Cj}^2 h_{jn} = [W_M B_C H_M]_{fn}, \tag{9}$$

where we introduced the following matrices: the $F \times J$ dictionary matrix W_M, such that $W_M(f,j) = w_{fj}$, the $J \times J$ matrix $B_C = \text{diag}(\beta_{Cj}^2)$ and the $J \times N$ amplitude coefficient matrix H_M, $H_M(j,n) = h_{jn}$.

3. METHOD

3.1 Maximum Likelihood (ML) parameter estimation

The parameters to be estimated are $\Theta = \{H_\Gamma, H_\Phi, H_{F_0}, W_M, H_M, \alpha_R, \alpha_L, B_R, B_L\}$. The dictionary matrices W_Γ and W_{F_0} are fixed:

- W_{F_0} is set using the glottal source model *KL-GLOTT88* [12] to obtain spectral "combs" plus, when needed, the unvoicing basis vector,
- W_Γ is set by using overlapping Hann functions covering the whole frequency range.

Thanks to our independency assumptions in time, frequency and between the channels, the log-likelihood $C(\Theta)$ of the observations writes:

$$C(\Theta) = \sum_{fn} \log N_c(X_{R,fn}; 0, S_{X_R,fn}) + \log N_c(X_{L,fn}; 0, S_{X_L,fn}), \quad (10)$$

where the variances for the left and right channels are given thanks to equations (4), (8) and (9):

$$S_{X_R,fn} = \alpha_R^2 \left[(W_\Gamma H_\Gamma H_\Phi).(W_{F_0} H_{F_0}) \right]_{fn} + \left[W_M B_R H_M \right]_{fn} \quad (11)$$

$$S_{X_L,fn} = \alpha_L^2 \left[(W_\Gamma H_\Gamma H_\Phi).(W_{F_0} H_{F_0}) \right]_{fn} + \left[W_M B_L H_M \right]_{fn} \quad (12)$$

3.2 Indeterminacies

The criterion (10) suffers from several indeterminacies. Scale indeterminacy arises first with the distribution of the energy between the dictionary matrices W_Γ, W_{F_0} and W_M and their corresponding amplitude matrices H_Γ (and H_Φ), H_{F_0} and H_M, as well as between H_Φ and H_{F_0}: we solve this problem by normalizing the columns of W_Γ, W_{F_0}, W_M, H_Γ and H_Φ. The scale indeterminacy concerning the mixing coefficients can be similarly avoided by normalizing the columns of the mixing matrix A.

3.3 Parameter updating rules

Our formalism is similar to Non-negative Matrix Factorization (NMF), especially for the accompaniment part as can be seen in [13]. We therefore derive an estimation algorithm that is based on "classical" derivations for this type of problems, and obtain multiplicative updating rules that iteratively aim at increasing our ML criterion (10) by re-estimating the parameters.

The partial derivatives for each parameter θ in Θ can always be represented in our case as: $\frac{\partial C(\Theta)}{\partial \theta} = P(\theta) - Q(\theta)$, where P and Q are positive functions. Following NMF methodology as in [4], the multiplicative updates are such that $\theta_{\text{new}} \leftarrow \theta_{\text{old}} \frac{P(\theta)}{Q(\theta)}$.

Algorithm 1 gives the resulting updating rules, where, for convenience, the power spectrograms are noted $D_{X_R} = |X_R|^2$ and $D_{X_L} = |X_L|^2$. S_{X_R} and S_{X_L} are computed thanks to equations (11) and (12). The solo parts S_{F_0} and S_Φ are computed with equations (6) and (7). We use a point '.' to represent Hadamard product and the convention for the power, ".(ω)", is used for elementwise exponentiation on the matrix. The "sum" operator stands for a summation over all the elements of the argument matrix.

There is no proof of convergence for Algorithm 1 but in our experiments, the criterion was always increasing over the iterations. The updating rules are very similar to "classical" NMF rules for Itakura-Saito divergence [13]. The normalizations discussed in section 3.2 are done after each update of the corresponding parameter. The matrices B_C are initialized with 0 for the off-diagonal elements, hence the updates (19) also return the desired diagonal matrices.

3.4 Solo/Accompaniment Separation System

The proposed iterative system flow is similar to [4]:

1. *1st (unconstrained) parameter estimation round* using Algorithm 1,

Algorithm 1 "IMM" Algorithm: Parameter updating rules for stereophonic signals

$$H_{F_0} \leftarrow H_{F_0}.\frac{W_{F_0}^T(\alpha_R^2 S_\Phi.S_{X_R}^{\cdot(-2)}.D_{X_R} + \alpha_L^2 S_\Phi.S_{X_L}^{\cdot(-2)}.D_{X_L})}{W_{F_0}^T(\alpha_R^2 S_\Phi.S_{X_R}^{\cdot(-1)} + \alpha_L^2 S_\Phi.S_{X_L}^{\cdot(-1)})} \quad (13)$$

$$H_\Phi \leftarrow H_\Phi.\frac{(W_\Gamma H_\Gamma)^T(\alpha_R^2 S_{F_0}.S_{X_R}^{\cdot(-2)}.D_{X_R} + \alpha_L^2 S_{F_0}.S_{X_L}^{\cdot(-2)}.D_{X_L})}{(W_\Gamma H_\Gamma)^T(\alpha_R^2 S_{F_0}.S_{X_R}^{\cdot(-1)} + \alpha_L^2 S_{F_0}.S_{X_L}^{\cdot(-1)})} \quad (14)$$

$$H_M \leftarrow H_M.\frac{(W_M B_R)^T(S_{X_R}^{\cdot(-2)}.D_{X_R}) + (W_M B_L)^T(S_{X_L}^{\cdot(-2)}.D_{X_L})}{(W_M B_R)^T(S_{X_R}^{\cdot(-1)}) + (W_M B_L)^T(S_{X_L}^{\cdot(-1)})} \quad (15)$$

$$H_\Gamma \leftarrow H_\Gamma.\frac{W_\Gamma^T(\alpha_R S_{F_0}.S_{X_R}^{\cdot(-2)}.D_{X_R} + \alpha_L S_{F_0}.S_{X_L}^{\cdot(-2)}.D_{X_L})H_\Phi^T}{W_\Gamma^T(\alpha_R S_{F_0}.S_{X_R}^{\cdot(-1)} + \alpha_L S_{F_0}.S_{X_L}^{\cdot(-1)})H_\Phi^T} \quad (16)$$

$$W_M \leftarrow W_M.\frac{(S_{X_R}^{\cdot(-2)}.D_{X_R})(B_R H_M)^T + (S_{X_L}^{\cdot(-2)}.D_{X_L})(B_L H_M)^T}{(S_{X_R}^{\cdot(-1)})(B_R H_M)^T + (S_{X_L}^{\cdot(-1)})(B_L H_M)^T} \quad (17)$$

$$\alpha_C \leftarrow \alpha_C \frac{\text{sum}\left(S_V.S_{X_C}^{\cdot(-2)}.D_{X_C}\right)}{\text{sum}\left(S_V.S_{X_C}^{\cdot(-1)}\right)}, \text{ for } C \in \{R, L\} \quad (18)$$

$$B_C \leftarrow B_C.\frac{W_M^T(S_{X_C}^{\cdot(-2)}.D_{X_C})H_M^T}{W_M^T(S_{X_C}^{\cdot(-1)})H_M^T}, \text{ for } C \in \{R, L\} \quad (19)$$

2. *Melody tracking:* a smooth path of fundamental frequencies is computed from the corresponding activation coefficients H_{F_0}, using a Viterbi algorithm; the chosen path thus accomplishes a trade-off between its energy and the transitions between successive f_0 frequencies,
3. *2nd parameter estimation round* using Algorithm 1 and $\bar{H}_{F_0}$ as initialisation for H_{F_0}:
   - Solo and accompaniment separation using the corresponding Wiener filters, on each channel
     → "voiced"-IMM (V-IMM),
4. *3rd parameter estimation round*, including the "unvoicing" basis vector in W_{F_0} and with W_Φ (i.e. H_Γ) fixed:
   - Separation by Wiener filters, on each channel
     → "voiced+unvoiced"-IMM (VU-IMM).

Figure 1 also depicts the flow of the system. Each of the three rounds of parameter estimation correspond to 500 iterations of Algorithm 1. For the first round, the parameters are randomly initialized with a set Θ_0. For the second round, they are also randomly initialized, except the amplitude matrix for the solo source part which is initialized as in [4]: a matrix $\bar{H}_{F_0}$ is obtained from the tracked main melody and the firstly estimated matrix H_{F_0} by setting to 0 all the coefficients that are outside a scope of a quarter tone from the estimated melody. These values remain null through the multiplicative rule (13). $\bar{H}_{F_0}$ is then used as initial H_{F_0} matrix for the second estimation round. After this second round, we obtain a first solo/accompaniment separation result (V-IMM), where only the voiced parts of the solo were taken into account. We obtain stereophonic STFT "images" of the estimated solo $\hat{V}_{\text{V-IMM}}$ and accompaniment $\hat{M}_{\text{V-IMM}}$ by applying the corresponding Wiener filters as in [4], individually on each channel. These images are such that:

$$\hat{V}_{\text{imag},fn} = \begin{bmatrix} \alpha_R \hat{V}_{R,fn} \\ \alpha_L \hat{V}_{L,fn} \end{bmatrix} \text{ and } \hat{M}_{\text{imag},fn} = \begin{bmatrix} \hat{M}_{R,fn} \\ \hat{M}_{L,fn} \end{bmatrix},$$

where $\alpha_R \hat{V}_{R,fn}$, $\alpha_L \hat{V}_{L,fn}$, $\hat{M}_{R,fn}$ and $\hat{M}_{L,fn}$ respectively are the Wiener estimators of the right and left channels of the solo and of

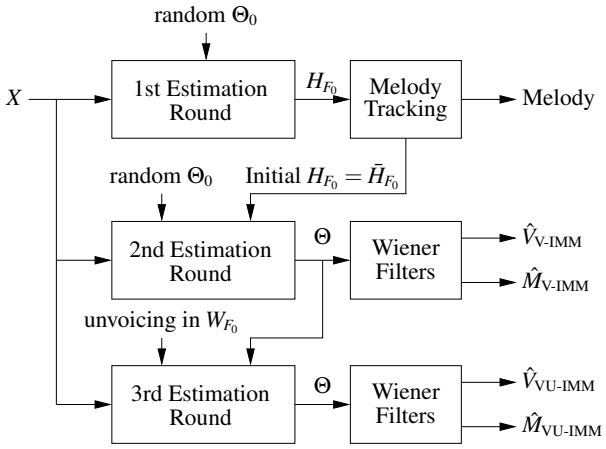

Figure 1: Solo/Accompaniment Separation System Flow

the right and left channels of the accompaniment. The audio tracks are then obtained by an overlap-add procedure applied individually on each channel of these STFT images.

At last, the initial parameters for the third round are the parameters estimated from the second round, except for W_{F_0}, to which we add the unvoiced basis vector. We assume, by fixing the filter dictionary W_Φ for this round, that the unvoiced parts of the solo instrument are generated by the same filters as for the voiced parts. The algorithm therefore catches unvoiced components whose spectral characteristics actually fit the previously estimated filter shapes. This new separation result is referred to as the VU-IMM, with the estimated images $\hat{V}_{\text{VU-IMM}}$ and $\hat{M}_{\text{VU-IMM}}$.

4. RESULTS

4.1 Database and evaluation criteria

Our development database is based on the multiple track recordings from the MTG MASS database [14]. We generated 13 synthetic instantaneous mixtures from the available multi-track data. The sampling rate is 44100Hz. The STFTs are computed on analysis windows of 2048 samples (46.44ms), with a hopsize of 256 samples (5.8ms, overlapping rate of 87.5%). We use a "sinebell" weighting window, both for analysis and synthesis.

The fundamental frequencies of the source part for the solo voice range from 60Hz to 1000Hz, with 96 f_0s per octave. This results in $N_{F_0} = 391$ basis vectors in W_{F_0}. For the filters, we chose $P = 30$ Hann atomic elements for W_Γ, with an overlap rate of 75%, covering the whole frequency range. This corresponds to elementary smooth filters with a constant "bandwidth" of about 3kHz. The number of filters is fixed to $K = 9$ and the number of spectra for the accompaniment to $R = 50$.

We evaluate our systems with the criteria defined for SiSEC, explained in [15]: Signal to Distortion Ratio (SDR), Image to Spatial distortion Ratio (ISR), Source to Interference Ratio (SIR) and Sources to Artefacts Ratio (SAR). We also refer to SDR (resp. SIR) "gains" (gSDR and gSIR) as being the difference between the SDR (resp. SIR) obtained by the estimated tracks and the SDR (resp. SIR) computed by setting the original stereophonic mixture as the estimated track.

We report the results for 5 systems in table 1: V-IMM and VU-IMM, each without (0) and with (1) the smooth filter model. The system "Mono" is the monophonic system [4], applied separately to each channel. Some audio examples are also available on-line[1].

[1] http://perso.telecom-paristech.fr/durrieu/en/eusipco09/

4.2 Smooth filters and unvoicing model

The results in table 1 first show that the performances in source separation with and without the smooth filter algorithm are not significantly different. This feature does not seem to be able to discriminate the timbre of the solo instrument from the accompaniment ones, since the extracted solo occasionally switches from the desired instrument to some instruments of the accompaniment. The main interest of obtaining smooth filters however lies in the better "semantics" that these spectral shapes may convey, rather than the direct improvement in source separation. As part of a production model for the solo instrument, the smoothness of the filters is more realistic than having unconstrained filters. It may therefore be useful for recognition purposes. In a supervised framework, it can also be used to learn spectral shapes that are characteristic for a given instrument.

The unvoicing model seems to lead to better results since VU-IMM in general obtained better results than V-IMM on our database. However, the difference between the criteria is not significantly high, and informal listening of the estimated tracks reveals that most of the unvoiced parts that are caught actually correspond to drum sounds. We also noticed that only some of the desired unvoiced solo parts are extracted: especially for the excerpts by "Tamy", from the SiSEC "Professionally Produced Music Recordings" [8], with a guitar as accompaniment instrument, some consonnants are missing in the extracted solo. This may show that the unvoicing model, i.e. assuming that the unvoiced parts share the same filter shapes as the voiced parts, is not complete and may need to be further extended in order to take into account the other potential unvoiced components.

4.3 Stereophonic vs. monophonic algorithm

In order to compare the monophonic algorithm [4] and the proposed algorithm in the same conditions, we create a "pseudo"-stereophonic result from [4] by applying the algorithm on both channels, separately and independently.

The table 1 shows that the performances are significantly improved by the use of the proposed stereophonic algorithms. In the stereophonic framework, the melody is estimated once for both channels and the energy variation for a single contribution, e.g. the solo, of one channel is therefore proportional to that of the other channel: the result is therefore more coherent, and this way we avoid obtaining separated signals that are randomly "floating" from one side to the other. Applying the monophonic algorithm independently on each channel does not guarantee this coherence.

Contrary to [3], our approach is also general enough to deal with "truly" stereophonic signals, even when the solo instrument is not exactly panned in the middle: allowing this flexibility therefore improves the separation of the accompaniment.

4.4 SiSEC campaign results

At last, early versions of the proposed systems were evaluated at the SiSEC evaluation campaign for "Professionally Produced Music Recordings" [8]. We provided the extracted female singer voice and the extracted background music (the guitar) for the first song by "Tamy" using two of the afore-mentionned methods, V-IMM and VU-IMM, both with smoothed filters.

The results in terms of SDR are given in table 2. Details for the other systems can be found in [15]. We ordered the systems by decreasing mean between the obtained SDR for the singer and guitar extractions. The result from the previous sections are confirmed, since VU-IMM performs better than V-IMM. Compared to the other participants, VU-IMM achieved the second mean SDR value, after Cancela [16], whose systems share an interesting similarity with our algorithms: they also explicitly model the solo part using the melodic line (with the fundamental frequencies in Hz). The results obtained during this evaluation tend to prove or at least to validate the use of this information in order to successfully separate (monophonic) melodic instruments. This type of "knowledge-based" approaches are then to be compared with more classical approaches for source separation, more "data-driven". The drawback is that

Method	SDR	ISR	SIR	SAR	gSDR	gSIR
Mono	5.8/6.9	9.0/21.8	16.8/9.6	5.8/11.5	6.9/5.8	17.8/8.5
V-IMM0	7.9/8.9	12.1/23.0	19.2/12.6	8.2/12.5	8.9/7.9	20.2/11.6
V-IMM1	7.9/8.9	12.5/22.1	18.4/12.8	8.3/11.6	8.9/7.9	19.4/11.8
VU-IMM0	**8.2/9.3**	12.4/**23.3**	**19.9**/12.9	8.7/**12.7**	**9.3/8.2**	**20.9**/11.8
VU-IMM1	**8.2/9.3**	**13.0**/21.8	18.6/**13.2**	**8.8**/12.0	**9.3/8.2**	19.6/**12.2**

Table 1: Average results on our database, in dB. For each criterion: estimated solo/estimated accompaniment

System	Singer SDR	Guitar SDR
Cancela2	**9.7**	8.6
VU-IMM	7.8	**9.4**
Cancela1	8.7	8.0
V-IMM	6.9	8.6
Cobos	6.4	8.0
Ozerov	5.1	6.7
Ozerov/Févotte	3.6	5.3
Vinyes Raso	4.9	4.2
Ideal Binary Mask	*10.1*	*11.8*

Table 2: Result table for SiSEC 2008 (song "Tamy - Que Pena / Tanto Faz")

the assumptions on the signals are quite strong, and it is difficult to adapt the model in order to extract some other instrument such as the guitar, for instance, from the other song by "Bearlin" in the SiSEC evaluation: such polyphonic instruments need a more complicated polyphonic pitch estimation step followed by a clustering step to determine which instrument played which estimated pitch.

5. CONCLUSION

The proposed algorithms extend our monophonic recording de-soloing scheme [4] to deal with stereophonic recordings. In spite of a rather simple model for stereophony, the results are satisfying and show again that our approach, which consists in estimating the melodic line to separate a monopitch instrument, especially a singer's voice, is adequate for de-soloing and solo enhancement purposes. In addition to stereophony, we proposed two other contributions in the source/filter model of the solo instrument: an explicit spectral smoothness of the filter part and an unvoicing model.
Even if the smoothness of the filters does not lead to better results, it is better in the context of a production model. It enables further studies such as song-to-text and lyrics-to-audio alignement. The unvoiced part that is added in the 3rd estimation round showed to be close to what was expected, but in presence of other strong "noisy" elements such as the drum sounds, it seems not to be discriminative enough. However, the resulting separated audio tracks seem, according to informal listening, to show that these unvoiced parts can be important in order to obtain an "intelligible" solo separation, when the solo instrument is a singer. Further studies on this model should aim at assessing the importance of including the unvoiced part to singing voice enhancement applications.
In order to further improve the stereophonic separation, the proposed signal model can be included into the more complete framework of [9], within which it could take advantage of more of the available mutual information between the channels.

REFERENCES

[1] E. Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech and Language Processing*, 14:91–98, 2006.

[2] O. Gillet and G. Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16:529–540, 2008.

[3] M. Ryynanen, 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*, pages 1417–1420, 2008.

[4] 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*, pages 105 – 108, 2009.

[5] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[6] D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience*, 2008.

[7] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 14:191–199, 2006.

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

[9] A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures. with application to blind audio source separation. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[10] E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 109 – 112, 2008.

[11] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech and Language Processing*, 15:1564–1578, 2007.

[12] D.H. Klatt and L.C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *JASA*, 1990.

[13] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Computation*, 21, 2009.

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

[15] E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *Proc. of International Conference on Independent Component Analysis and Signal Separation*, 2009.

[16] P. Cancela. Tracking melody in polyphonic audio. mirex 2008. *Proc. of Music Information Retrieval Evaluation eXchange*, 2008.