MULTICHANNEL HARMONIC AND PERCUSSIVE COMPONENT SEPARATION BY JOINT MODELING OF SPATIAL AND SPECTRAL CONTINUITY

Ngoc Q. K. Duong[1]*, Hideyuki Tachibana*[2]*, Emmanuel Vincent*[1]
Nobutaka Ono[2]*, Rémi Gribonval*[1] *and Shigeki Sagayama*[2]

[1] INRIA, Centre de Rennes - Bretagne Atlantique, France
qduong@irisa.fr, emmanuel.vincent@inria.fr, remi.gribonval@inria.fr
[2] Department of Information Physics and Computing,
Graduate School of Information Science and Technology, The University of Tokyo
{tachibana,onono,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

This paper considers the blind separation of the harmonic and percussive components of multichannel music signals. We model the contribution of each source to all mixture channels in the time-frequency domain via a *spatial covariance* matrix, which encodes its spatial characteristics, and a scalar *spectral variance*, which represents its spectral structure. We then exploit the spatial continuity and the different spectral continuity structures of harmonic and percussive components as prior information to derive maximum a posteriori (MAP) estimates of the parameters using the expectation-maximization (EM) algorithm. Experimental results over professional musical mixtures show the effectiveness of the proposed approach.

Index Terms— Harmonic and percussive source separation, local Gaussian model, continuity prior.

1. INTRODUCTION

Sounds in most real-world musical mixtures can be classified into two major types depending on their spectral structure: harmonic sounds such as vocals, piano or violin and percussive sounds such as drums. Denoting by $\mathbf{h}(n,f)$ and $\mathbf{p}(n,f)$ the contribution of harmonic sounds and percussive sounds to all I mixture channels in the short time Fourier transform (STFT) domain in time frame n and frequency bin f, respectively, the observed $I \times 1$ mixture signal $\mathbf{x}(n,f)$ can be expressed as

$$\mathbf{x}(n,f) = \mathbf{h}(n,f) + \mathbf{p}(n,f). \tag{1}$$

The separation of these components from a mixture is useful for remixing [1] and also for various music information retrieval tasks [2–4]. In [1] a method called Harmonic/Percussive Sound Separation (HPSS) was proposed whereby the harmonic and percussive sources can be blindly separated in the time-frequency domain from a single channel mixture using the assumption that spectrograms of harmonic components are *smooth in the time direction* while those of percussive components are *smooth in the frequency direction*. However, this technique does not exploit the spatial information available in multichannel mixtures and does not readily extend to this context due to the chosen divergence cost.

In this paper, we address the separation of harmonic and percussive components from multichannel mixtures using the local Gaussian modeling framework in [5]. We model $\mathbf{h}(n,f)$ and $\mathbf{p}(n,f)$ as zero-mean Gaussian random variables whose respective covariances $\mathbf{\Sigma}_h(n,f)$ and $\mathbf{\Sigma}_p(n,f)$ are factorized into

$$\begin{aligned} \mathbf{\Sigma}_h(n,f) &= v_h(n,f)\mathbf{R}_h(n,f) \\ \mathbf{\Sigma}_p(n,f) &= v_p(n,f)\mathbf{R}_p(n,f) \end{aligned} \tag{2}$$

where $v_h(n,f)$ and $v_p(n,f)$ are scalar time-varying *spectral variances* encoding the spectro-temporal power of harmonic and percussive components, respectively, while $\mathbf{R}_h(n,f)$ and $\mathbf{R}_p(n,f)$ are $I \times I$ full-rank time-varying *spatial covariance matrices* encoding their spatial position and spatial spread. Note that, compared to [5], we do not assume that the spatial covariance matrices are constant over time. Rather, we introduce a continuity prior for the spatial covariance matrices and incorporate the temporal and spectral smoothness objectives in [1] using different priors for the spectral variances of harmonic and percussive components. The parameters are then estimated in the maximum a posteriori (MAP) sense. Finally, the separated components are obtained in the minimum mean square error (MMSE) sense by multichannel Wiener filtering. Matlab code for the proposed algorithm is available [1].

The structure of the rest of the paper is as follows. We present the model of spatial and spectral continuity in Section 2 and address MAP estimation of the model parameters in Section 3. We provide experimental results to confirm the effectiveness of the proposed approach in Section 4 and finally we conclude in Section 5.

[1] https://www.irisa.fr/metiss/ngoc/sw/hpss.rar

2. SPATIAL AND SPECTRAL CONTINUITY MODELING

Under the mixing model (1) and the parameterization (2), assuming that the components are uncorrelated, the vector of STFT coefficients of the mixture signal $\mathbf{x}(n,f)$ is zero-mean Gaussian with covariance matrix

$$\mathbf{\Sigma}_x(n,f) = v_h(n,f)\mathbf{R}_h(n,f) + v_p(n,f)\mathbf{R}_p(n,f). \quad (3)$$

The log-likelihood is then given by

$$\log \mathcal{L} = -\sum_{n,f} \mathrm{tr}(\mathbf{\Sigma}_x^{-1}(n,f)\widehat{\mathbf{\Sigma}}_x(n,f)) + \log\det(\pi\,\mathbf{\Sigma}_x(n,f)) \quad (4)$$

where $\det(.)$ denotes the determinant of a square matrix and $\widehat{\mathbf{\Sigma}}_x(n,f)$ the empirical mixture covariance matrix as defined in [6].

We enforce spatial and spectral smoothness of harmonic and percussive components by introducing prior distributions for $\mathbf{R}_h(n,f), \mathbf{R}_p(n,f)$, $v_h(n,f)$, and $v_p(n,f)$, and estimating them in the MAP sense.

2.1. Spatial continuity prior

When each component consists of a single harmonic or percussive source and reverberation is moderate, the spatial covariance matrices are time-invariant and may be modeled as in [5]. However, in general, each component consists of several sources, *e.g* the percussive component includes several drums (bass drum, snare drum, hi-hat, etc). The spatial covariance matrices of each component are then time-varying but can be assumed to vary smoothly over time due to the fact that one source usually predominates in a given time-frequency neighborhood. We then choose the following Markov chain prior for $\mathbf{R}_h(n,f)$ and $\mathbf{R}_p(n,f)$ when $n > 1$

$$\begin{aligned} p\big(\mathbf{R}_h(n,f)\big) &= \mathcal{IW}\big(\mathbf{R}_h(n,f)|(m_h - I)\mathbf{R}_h(n-1,f), m_h\big) \\ p\big(\mathbf{R}_p(n,f)\big) &= \mathcal{IW}\big(\mathbf{R}_p(n,f)|(m_p - I)\mathbf{R}_p(n-1,f), m_p\big). \end{aligned} \quad (5)$$

$\mathcal{IW}\big(\mathbf{R}|\mathbf{\Psi}, m\big)$ denotes the inverse Wishart density over positive definite matrices $\mathbf{R}$ with positive definite inverse scale matrix $\mathbf{\Psi}$ and m degrees of freedom [7]

$$\mathcal{IW}(\mathbf{R}|\mathbf{\Psi}, m) = \frac{|\mathbf{\Psi}|^m |\mathbf{R}|^{-(m+I)} e^{-\mathrm{tr}(\mathbf{\Psi}\mathbf{R}^{-1})}}{\pi^{I(I-1)/2}\prod_{i=1}^{I}\Gamma(m-i+1)} \quad (6)$$

with $\mathrm{tr}(.)$ denoting the trace of a square matrix and Γ the Gamma function such that the mean of $\mathbf{R}$ is given by $\mathbf{\Psi}/(m-I)$ [7]. This distribution, its mean, and its variance exists for $m > I-1$, $m > I$, and $m > I+1$ respectively

The reason for which we consider an inverse-Wishart prior for the spatial covariance matrices is that it is the conjugate prior for the likelihood of the considered Gaussian observation model, which results in closed-form updates. The initial distributions $p\big(\mathbf{R}_h(1,f)\big)$ and $p\big(\mathbf{R}_p(1,f)\big)$ are chosen as uniform for all f.

2.2. Spectral continuity prior

Since the spectrum of harmonic components is usually *smooth* over the time axis while that of percussive components is usually *smooth* over the frequency axis [1], we consider the following Markov chain priors for $v_h(n,f)$ with $n > 1$ and $v_p(n,f)$ with $f > 1$:

$$\begin{aligned} p\big(v_h(n,f)\big) &= \mathcal{IG}\big(v_h(n,f)|\alpha_h, (\alpha_h - 1)v_h(n-1,f)\big) \\ p\big(v_p(n,f)\big) &= \mathcal{IG}\big(v_p(n,f)|\alpha_p, (\alpha_p - 1)v_p(n,f-1)\big) \end{aligned} \quad (7)$$

$\mathcal{IG}(v|\alpha,\beta)$ denotes the inverse-gamma density with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$

$$\mathcal{IG}(v|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} v^{-\alpha-1} e^{-\beta/v} \quad (8)$$

whose mean is $\beta/(\alpha-1)$.

Similarly to the choice of initial distribution for the spatial covariances, $p\big(v_h(1,f)\big)\ \forall f$ and $p\big(v_p(n,1)\big)\ \forall n$ are chosen as uniform. The choice of an inverse-gamma prior, which is the conjugate prior for the considered likelihood, brings not only simpler computation compared to the Gaussian prior in [1] but also better separation performance as shown in our experiments in Section 4. This prior distribution was also used to model temporal continuity in [8] in the different context of multipitch estimation.

3. MAP ESTIMATION OF MODEL PARAMETERS

We derive an EM algorithm to estimate the spatial parameters $\theta_{\text{spat}} = \{\mathbf{R}_h(n,f), \mathbf{R}_p(n,f)\}_{n,f}$ and the spectral parameters $\theta_{\text{spec}} = \{v_h(n,f), v_p(n,f)\}_{n,f}$ of the two types of components in each time-frequency bin (n,f) using the *complete data* $\mathbf{c} = \{\mathbf{h}(n,f), \mathbf{p}(n,f)\}_{n,f}$, that is the set of STFT coefficients of the harmonic component and the percussive component in all time-frequency bins.

In the E-step, the expected covariance matrices $\widehat{\mathbf{\Sigma}}_h(n,f)$, $\widehat{\mathbf{\Sigma}}_p(n,f)$ are updated similarly as in [6] using the Wiener filters $\mathbf{W}_h(n,f)$ and $\mathbf{W}_p(n,f)$

$$\mathbf{W}_h(n,f) = \mathbf{\Sigma}_h(n,f)\mathbf{\Sigma}_x^{-1}(n,f) \quad (9)$$

$$\mathbf{W}_p(n,f) = \mathbf{\Sigma}_p(n,f)\mathbf{\Sigma}_x^{-1}(n,f) \quad (10)$$

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_h(n,f) = \mathbf{W}_h(n,f)\widehat{\mathbf{\Sigma}}_x(n,f)\mathbf{W}_h^H(n,f) \\ + (\mathbf{I} - \mathbf{W}_h(n,f))\mathbf{\Sigma}_h(n,f) \end{aligned} \quad (11)$$

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_p(n,f) = \mathbf{W}_p(n,f)\widehat{\mathbf{\Sigma}}_x(n,f)\mathbf{W}_p^H(n,f) \\ + (\mathbf{I} - \mathbf{W}_p(n,f))\mathbf{\Sigma}_p(n,f) \end{aligned} \quad (12)$$

where $\mathbf{I}$ is the $I \times I$ identity matrix, $\mathbf{\Sigma}_h(n,f)$, $\mathbf{\Sigma}_p(n,f)$ are defined in (2), $\mathbf{\Sigma}_x(n,f)$ in (3), and $\widehat{\mathbf{\Sigma}}_x(n,f)$ in [6].

In the M-step, the auxiliary function Q defined in the MAP sense as

$$\begin{aligned} Q_{MAP}(\theta|\theta^{\text{old}}) \quad &= \log p\big(\mathbf{c}|\theta\big) + \gamma_1 \log p\big(\theta_{\text{spat}}\big) \\ &+ \gamma_2 \log p\big(\theta_{\text{spec}}\big) \end{aligned} \quad (13)$$

is maximized with respect to the parameters $\theta = \{\theta_{spat}, \theta_{spec}\}$, where γ_1, γ_2 are tradeoff hyper-parameters determining the contribution of the priors and

$$p\big(\mathbf{c}|\theta\big) = \prod_{n,f} p\big(\mathbf{h}(n,f)|\mathbf{0}, \mathbf{\Sigma}_h(n,f)\big) p\big(\mathbf{p}(n,f)|\mathbf{0}, \mathbf{\Sigma}_p(n,f)\big) \tag{14}$$

$$p\big(\theta_{\text{spat}}\big) = \prod_{n,f} p\big(\mathbf{R}_h(n,f)\big) p\big(\mathbf{R}_p(n,f)\big) \tag{15}$$

$$p\big(\theta_{\text{spec}}\big) = \prod_{n,f} p\big(v_h(n,f)\big) p\big(v_p(n,f)\big). \tag{16}$$

By substituting (5) into (15), (7) into (16), and (14), (15), (16) into (13), then computing the gradient of $Q_{MAP}(\theta|\theta^{\text{old}})$ with respect to each entry of $\mathbf{R}_h(n,f)$, $\mathbf{R}_p(n,f)$ and equating it to zero, we obtain a quadratic matrix equation. After solving this equation under the constraint that the solution is positive definite [9], the spatial covariance matrices are updated as

$$\begin{aligned}\mathbf{R}_h(n,f) &= (1/2)\mathbf{A}_h^{-1}\big(-\mathbf{B} + (\mathbf{B}^2 - 4\mathbf{A}_h\mathbf{C}_h\mathbf{A}_h)^{1/2}\big)\mathbf{A}_h^{-1} \\ \mathbf{R}_p(n,f) &= (1/2)\mathbf{A}_p^{-1}\big(-\mathbf{B} + (\mathbf{B}^2 - 4\mathbf{A}_p\mathbf{C}_p\mathbf{A}_p)^{1/2}\big)\mathbf{A}_p^{-1}\end{aligned} \tag{17}$$

where $(.)^{1/2}$ denotes the square root of a Hermitian matrix, and

$$\begin{aligned}\mathbf{A}_h &= \big(\gamma_1(m_h - I)\mathbf{R}_h^{-1}(n+1,f)\big)^{1/2} \\ \mathbf{A}_p &= \big(\gamma_1(m_p - I)\mathbf{R}_p^{-1}(n+1,f)\big)^{1/2} \\ \mathbf{B} &= (\gamma_1 I + 1)\mathbf{I} \\ \mathbf{C}_h &= -\widehat{\mathbf{\Sigma}}_h(n,f)/v_h(n,f) - \gamma_1(m_h - I)\mathbf{R}_h(n-1,f) \\ \mathbf{C}_p &= -\widehat{\mathbf{\Sigma}}_p(n,f)/v_p(n,f) - \gamma_1(m_p - I)\mathbf{R}_p(n-1,f)\end{aligned} \tag{18}$$

Similarly, by computing the gradient of $Q_{MAP}(\theta|\theta^{\text{old}})$ with respect to $v_h(n,f)$, $v_p(n,f)$ and equating it to zero, we get a second order polynomial form of the source variances with a single positive solution, that is

$$\begin{aligned}v_h(n,f) &= (-b + \sqrt{b^2 - 4a_h c_h})/(2a_h) \\ v_p(n,f) &= (-b + \sqrt{b^2 - 4a_p c_p})/(2a_p)\end{aligned} \tag{19}$$

where

$$\begin{aligned}a_h &= \gamma_2(\alpha_h - 1)/v_h(n+1,f) \\ a_p &= \gamma_2(\alpha_p - 1)/v_p(n,f+1) \\ b &= \gamma_2 + I \\ c_h &= -\operatorname{tr}\big(\mathbf{R}_h^{-1}(n,f)\widehat{\mathbf{\Sigma}}_h(n,f)\big) - \gamma_2(\alpha_h - 1)v_h(n-1,f) \\ c_p &= -\operatorname{tr}\big(\mathbf{R}_p^{-1}(n,f)\widehat{\mathbf{\Sigma}}_p(n,f)\big) - \gamma_2(\alpha_p - 1)v_p(n,f-1)\end{aligned} \tag{20}$$

Note that, in (18) and (20), due to the choice of initial distributions $\mathbf{R}_h(n-1,f)$, $\mathbf{R}_p(n-1,f)$, $v_h(n-1,f)$ are zero for all f when $n = 1$, and $v_p(n,f-1)$ is zero for all n when $f = 1$. Source variances are uniformly initialized as $v_h(n,f) = v_p(n,f) = 1$ for all n, f while the spatial covariance matrices are initialized from the observed mixture covariance as $\mathbf{R}_h(n,f) = \mathbf{R}_p(n,f) = \frac{1}{2}\widehat{\mathbf{\Sigma}}_x(n,f)\ \forall n, f$. The algorithm converges just after five EM iterations. Finally, the separated components are obtained by multichannel Wiener filtering as

$$\begin{aligned}\widehat{\mathbf{h}}(n,f) &= \mathbf{W}_h(n,f)\mathbf{x}(n,f) \\ \widehat{\mathbf{p}}(n,f) &= \mathbf{W}_p(n,f)\mathbf{x}(n,f)\end{aligned} \tag{21}$$

4. EXPERIMENTAL RESULTS

Sound mixing techniques vary depending on the music genre: for certain genres, instruments are placed at different spatial positions while for other genres they are all mixed to the center. We evaluated the separation performance of the proposed algorithm over 8 stereo music mixtures of harmonic and percussive sources corresponding to two different mixing conditions. These mixtures are part of the Quaero project database, which was used for the *Professionally produced music recordings* task of the 2010 Signal Separation Evaluation Campaign (SiSEC 2010) [2]. The first 4 mixtures were originally mixed by a sound engineer where most instruments are panned close to the center with artificial reverb, and the total number of harmonic and percussive sources in each mixture varies from four to eight. In order to investigate the contribution of spatial information, in the second set of mixtures (named Pan+) we moved each source to a random position by amplitude panning but keeping the same reverb. The parameter setting is summarized in Table 1. In this experiment, the empirical mixture covariance $\widehat{\mathbf{\Sigma}}_x(n,f)$ was computed by local averaging as in [6] and the hyper-parameters m_h, m_p, α_h, α_p are heuristically fixed depending on the desired shape of the priors, which determines the degree of smoothness.

Mixture signal duration	10 s
Number of channels	$I = 2$
Sampling rate	44100 Hz
STFT frame size	4096
STFT frame shift	2048
Number of EM iterations	5
$\mathcal{IG}$ shape parameters	$\alpha_h = \alpha_p = 10$
$\mathcal{IW}$ degrees of freedom	$m_h = m_p = 5$
Trade-off parameters	$\gamma_1 = 0.5,\ \gamma_2 = 1$

Table 1. Common experimental parameter settings

Separation performance was evaluated using the widely used signal-to-distortion ratio (SDR) criterion measuring the overall distortion, as well as the signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) and source image-to-spatial distortion ratio (ISR) criteria in [10]. We compared the performance of the proposed multichannel har-

[2] http://sisec.wiki.irisa.fr/

monic and percussive sound separation algorithm (M-HPSS) with that achieved by the original single channel HPSS algorithm using I-divergence and Gaussian continuity priors (HPSS) [1], and with that given by the single channel HPSS algorithm introduced in this paper using inverse-gamma prior ($\text{HPSS}_{\mathcal{IG}}$). The results were averaged over all mixtures for each dataset and are shown in Table 2. All mixtures and the harmonic/percussive signals separated via the 3 tested methods are available on our webpage [3].

		SDR	SIR	SAR	ISR
Original	HPSS	3.8	5.2	7.6	8.7
	$\text{HPSS}_{\mathcal{IG}}$	4.8	7.9	8.0	10.7
	M-HPSS	5.0	7.2	8.6	10.1
Pan+	HPSS	3.8	5.1	7.5	8.6
	$\text{HPSS}_{\mathcal{IG}}$	4.7	7.7	8.2	10.4
	M-HPSS	5.3	7.4	8.8	10.3

Table 2. Average harmonic/percussive component separation performance

The numerical results show the significant separation improvement of $\text{HPSS}_{\mathcal{IG}}$ compared to the original HPSS in terms of all criteria over both datasets. This means that the inverse-gamma prior investigated in this paper better models the spectral continuity of harmonic and percussive components than the Gaussian prior introduced in [1]. Separation performance given by HPSS and $\text{HPSS}_{\mathcal{IG}}$ over the Pan+ dataset is very similar to that over the original dataset due to the fact that panning does not affect the spectral structure of the sources. But the performance achieved by M-HPSS has noticeably increased in panned datasets, *i.e.* the SDR is 0.6 dB higher than that given by $\text{HPSS}_{\mathcal{IG}}$ showing the benefit of exploiting spatial information. The performance improvements are confirmed by informal listening test and by the additional auditory-motivated criteria introduced at SiSEC 2010.

5. CONCLUSION

In this paper, we proposed a *multichannel* approach for the separation of harmonic and percussive components in musical recordings by joint modeling of spatial and spectral continuity. We investigated suitable continuity priors for both spatial and spectral parameters such that they are estimated in the MAP sense using the EM algorithm. Experimental results over professional musical mixtures confirm the benefit of the proposed multichannel approach compared to single-channel algorithms. Future work will investigate the learning of hyper-parameters and consider the proposed separation algorithm as a pre-processing step for some MIR tasks, *e.g.* singing voice extraction.

[3] https://www.irisa.fr/metiss/ngoc/sw/HPSSresults.rar

6. ACKNOWLEDGMENT

This work was supported by INRIA under the Associate Team Program VERSAMUS (http://versamus.inria.fr).

7. REFERENCES

[1] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. ISMIR*, 2008, pp. 139–144.

[2] J-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *Proc. ICASSP*, 2008, pp. 169–172.

[3] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, pp. 529–540, 2008.

[4] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody line estimation in homophonic music audio signal based on temporal-variability of melodic source," in *Proc. ICASSP*, 2010, pp. 425–428.

[5] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[6] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using local observed covariance and auditory-motivated time-frequency representation," in *Proc. LVA/ICA*, Sep. 2010, pp. 73–80.

[7] D. Maiwald and D. Kraus, "Calculation of moments of complex Wishart and complex inverse-Wishart distributed matrices," *on Radar, Sonar and Navigation, IEE Proceedings*, vol. 147, pp. 162–168, 2000.

[8] 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, pp. 793–830, 2009.

[9] N. J. Higham and H. M. Kim, "Solving a quadratic matrix equation by Newton's method with exact line searches," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, pp. 303–316, 2001.

[10] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results," in *Proc. ICA*, 2007, pp. 552–559.