MOSPALOSEP: A PLATFORM FOR THE BINAURAL LOCALIZATION AND SEPARATION OF SPATIAL SOUNDS USING MODELS OF INTERAURAL CUES AND MIXTURE MODELS

Joan Mouba[*]

ESIGETEL – LRIT
1 rue du Port de Valvins, Avon, 77210, France
joan.mouba@esigetel.fr

ABSTRACT

In this paper, we present the MOSPALOSEP platform for the localization and separation of binaural signals. Our methods use short-time spectra of the recorded binaural signals. Based on a parametric model of the binaural mix, we exploit the joint evaluation of interaural cues to derive the location of each time-frequency bin. Then we describe different approaches to establish localization: some based on an energy-weighted histogram in azimuth space, and others based on an unsupervised number of sources identification of Gaussian mixture model combined with the Minimum Description Length. In this way, we use the revealed Gaussian Mixture Model structure to identify the particular region dominated by each source in a multi-source mix. A bank of spatial masks allows the extraction of each source according to the posterior probability or to the Maximum Likelihood binary masks. An important condition is the Windowed-Disjoint Orthogonality of the sources in the time-frequency domain. We assess the source separation algorithms specifically on instruments mix, where this fundamental condition is not satisfied.

1. INTRODUCTION

In active listening, the separation of a stereo signal is a crucial pre-processing tool for the interaction with the individual sources which can be heard in the mix, for example by changing their spatial position. In fact, in a recording, the sound engineer may wish to change the position of the guitar, or to remove the singer's voice (for the karaoke effect), or remove all the instruments keeping only the lead voice (a cappella). It is a challenge to separate the sources in our case as there are only two sensors (miniature microphones embedded at human ears) and no restriction on the number of sources. Three early approaches are described in [1], [2], [3].

In [4], we propose a algorithm for the separation of an arbitrary number of audio sources where there is a binaural signal. This method uses a uni-dimensional power-weighted histogram constructed in the azimuth space and modeled as a Gaussian Mixture Model (GMM). The GMM structure *e.g.* number of sources, weight, azimuthal location and deviation of each source is calculated using a Maximum Likelihood (ML) approach based on an Expectation Maximization (EM) [5]. The GMM parameters are used to setup a source separation stage where the energy of each bin of the mix is assigned according to a posterior probability mask, but the number of sources is not identified automatically. The system's processing overview is depicted in Figure 1. In this paper, we compare the histogram based localization methods to an EM based localization method associated with the minimization of the Minimum Description Length (MDL), and we discuss the estimation of a precise number of sources; then we investigate the source separation performance for complex musical signals mix.

This paper is organized as follows. We start by a presentation of the binaural signal model in Section 2, then we describe the joint azimuth estimator and the localization methods in Section 3. In Section 4, we detail the source separation algorithms. Section 5 is dedicated to the description of our developed Toolbox, called MOSPALOSEP (system for MOdeling, SPAtialization, LOcalization and Separation). In Section 6, we conduct a comparative study of the localization algorithms, and we investigate the source separation performance with a probabilistic mask in comparison to a binary mask. Finally, we conclude and propose future works.

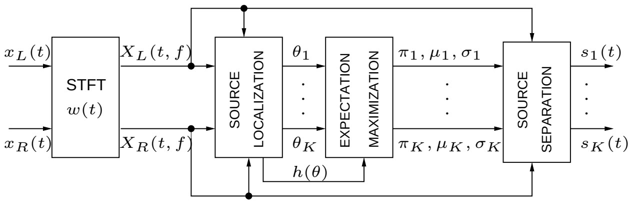


Figure 1: *Overview of the system's processing.*

2. MODEL

We considered punctual and omni-directional sound sources in the horizontal plane where both the listeners and the musicians are on the same level. Each source is located by its (ρ, θ) coordinates (no elevation), where ρ is the distance from the source to the listener head's center and θ is the azimuth angle.

In a binaural context, the difference in amplitude or Interaural Level Difference (ILD, expressed in decibels – dB) and in arrival time or Interaural Time Difference (ITD, expressed in seconds) are the main spatial cues for the auditory system [6]. In fact, a sound source positioned on the left will reach the left ear sooner than the right one, in the same manner the right amplitude level should be lower because of wave propagation and head shadowing.

These binaural cues can be related to physical parameters such as the speed of sound c and the head radius r. From the analysis of the CIPIC database, Viste [1] derives a sinusoidal model for the ILD. In [7], we also propose a sinusoidal model for the ITD. Both models are given with:

$$\mathrm{ILD}(\theta, f) = \alpha_f \sin(\theta), \quad (1)$$

$$\mathrm{ITD}(\theta, f) = \beta_f r \sin(\theta)/c, \quad (2)$$

[*] Part of this research took place at LaBRI – University of Bordeaux.

where α_f and β_f are frequency-dependent scaling factors (see [7]), that best fit for the models covering the 45 subjects of the CIPIC database for all azimuths, in a least-square sense.

For K spatially distributed sources $s_k(n)$ located at azimuth θ_k, the binaural signal mix can be expressed as:

$$\begin{aligned} x_L(n) &= \sum_{k=1}^{K} s_k, & (3) \\ x_R(n) &= \sum_{k=1}^{K} a_k(n) * s_k(n - d_k(n)), & (4) \end{aligned}$$

where $*$ is the convolution operator, $a_k(n)$ is the amplitude of the k-th source for the right ear microphone relative to the left ear microphone, $d_k(n)$ is the delay between the two microphones.

The convolution in the time-domain is equivalent to the multiplication in the frequency domain. By applying a Short-Time Fourier Transform on Equations (9) and (10), the model becomes:

$$\begin{aligned} X_L(t,f) &= \sum_{k=1}^{K} S_k(t,f), & (5) \\ X_R(t,f) &= \sum_{k=1}^{K} 10^{-\Delta_{a,k}(f)} S_k(t,f) \cdot e^{-j\Delta_{d,k}(f)}, & (6) \end{aligned}$$

where $\Delta_{a,k}(t,f)$ and $\Delta_{d,k}(t,f)$ are given by:

$$\begin{aligned} \Delta_{a,k}(f) &= \text{ILD}(\theta_k, f)/20, & (7) \\ \Delta_{d,k}(f) &= \text{ITD}(\theta_k, f) \cdot 2\pi f. & (8) \end{aligned}$$

Thus, to separate the sources $s_k(n)$ given the two channels $x_L(n)$ and $x_R(n)$, we need to detect the number of sources K and the mixing parameter θ_k for each source k.

3. SOURCE LOCALIZATION

The most important condition is that we consider any pair of sources $(s_k(t), s_l(t))$ as Windowed-Disjoint Orthogonal (WDO) [3]. This means that their short-time spectra do not overlap. Although, speech signals are approximately WDO, this condition is rarely satisfied for music signals. Based on this condition, the Equations (5) and (6) can be simplified to:

$$\begin{aligned} \Delta_{a,k}(f) &= \log_{10} \left| \frac{X_R(t,f)}{X_L(t,f)} \right|, & (9) \\ \Delta_{d,k}(f) &= \angle \frac{X_R(t,f)}{X_L(t,f)} + 2\pi p, & (10) \end{aligned}$$

where $\angle$ is the angle operator. The coefficient p shows that the phase can be determined up to a modulo 2π factor. In fact, the phase becomes ambiguous beyond 1500 Hz, according to the Duplex Theory.

Obtaining estimations of the azimuth based on the ILD and ITD information (for each p candidate) is simply a matter of inverting Equations (1) and (2).

The detected location θ is the one that minimizes the distance between θ_L and $\theta_{T,p}$, where

$$\begin{aligned} \theta_L(t,f) &= \arcsin\left(\frac{\text{ILD}(t,f)}{\alpha(f)}\right), & (11) \\ \theta_{T,p}(t,f) &= \arcsin\left(\frac{c \cdot \text{ITD}_p(t,f)}{r \cdot \beta(f)}\right). & (12) \end{aligned}$$

The $\theta_L(t,f)$ estimates are more dispersed, but not ambiguous at any frequency, so they are exploited as the azimuth reference. The optimization problem is equivalent to the finding of the right modulo coefficient p that unwraps the phase of the $\theta_{T,p}(t,f)$ estimate.

After Viste [1], a good solution is the $\theta_{T,p}(t,f)$ which is nearest to $\theta_L(t,f)$, as it exhibits a smaller deviation:

$$\theta(t,f) = \theta_{T,m}(t,f), \qquad (13)$$

$$\text{with } m = \text{argmin}_p \left| \theta_L(t,f) - \theta_{T,p}(t,f) \right| .$$

In practical terms, it is enough to restrict the choice of p among in the pair $(\lceil p_r \rceil, \lfloor p_r \rfloor)$, where

$$p_r = \left(f \cdot \text{ITD}(\theta_L, f) - \frac{1}{2\pi} \angle \frac{X_L(t,f)}{X_R(t,f)} \right). \qquad (14)$$

For each frequency bin of each discrete spectrum, an azimuth is then estimated.

3.1. Histogram based localization and identification of the number of sources

For each frequency bin's azimuth, we compute its corresponding power, and add it to the others in a spatial histogram (see [4]). The locations of the sources are obtained as the abscissa of the dominant peaks in the energy histogram. We leveled out the histogram and applied a threshold in order to refine the estimations, and to remove erroneous detected sources (see Figure 2). The number of local maxima is an estimate of the number of sources.

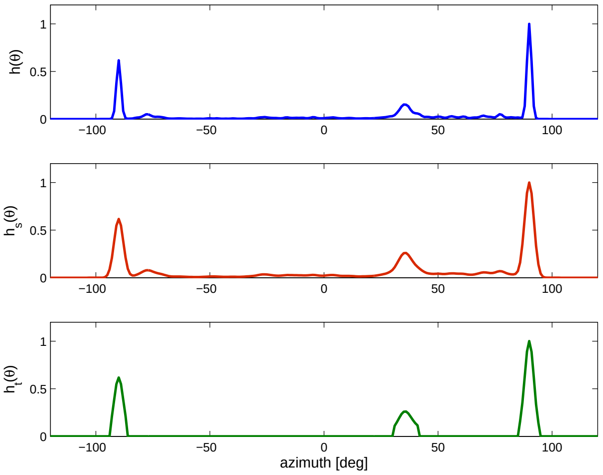


Figure 2: *Histograms derived from a mix of 3 speech sources at* $(-80°, +35°, +80°)$. *From top to bottom: raw histogram, leveled out histogram and capped histogram at* -40 *dB.*

3.2. EM with MDL based localization and number of sources identification

Here we collect the azimuth of each frequency bin, then we compute the underlying Gaussian Mixture Model with Expectation Maximization combined with the Minimum Description Length criterion, we obtain source locations and number of sources.

3.2.1. *Gaussian Mixture Model*

The sources are not exactly WDO, thus for each source we obtain a distribution around the true value. We characterize each source in the mix with a Gaussian representation. We then considered the mix as a K Gaussian (K-GMM):

$$P_K(\theta|\Gamma) = \sum_{k=1}^{K} \pi_k \ \phi_k(\theta|\mu_k, \sigma_k) \text{ with } \pi_k \geq 0 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \tag{15}$$

where Γ is a multiset of K triples $(\pi_k, \mu_k, \sigma_k^2)$ which denotes all the parameters of the model; π_k, μ_k, and σ_k^2 indicate respectively the weight, the mean, and the variance of the k-th Gaussian in:

$$\phi_k(\theta|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\theta-\mu_k)^2}{2\sigma_k^2}\right). \tag{16}$$

K represents the number of sources and Γ is the complete set of parameters. In one approach we use the EM combined with the penalty criterion MDL.

3.2.2. *Minimum Description Length*

The MDL is a criterion suggested by Rissanen [8] to find the optimal GMM structure and its number of sources by attempting to minimize the number of bits needed to code both the data θ and the parameter Γ. It is computed with:

$$MDL(K, \Gamma) = -\log P_K(\theta|\Gamma) + \frac{1}{2} L \log(NM) \tag{17}$$

with $L = K(1+M+\frac{(M+1)M}{M})$, N is the number of data samples and M is the dimension of the mean vector, which, in this case is one (one dimensional azimuth data).

The maximization of the MDL is a complex task, since for each number of sources K an EM algorithm must be run. By starting with a high value of K, the number of sources is reduced by merging one pair of clusters (u, v) that minimizes the distance $d(u, v)$ [9].

3.2.3. *Expectation Maximization*

Expectation Maximization is a well-known method to estimate parameters in mixture densities. The idea is to complete the observed data θ with an unobserved variable k to form the complete data (θ, k), where $k \in \{1, \cdots, K\}$ indicates the index of the Gaussian component from which θ has been drawn.

EM is an iterative algorithm, at each iteration we calculate the optimal parameters which increase the log-likelihood of the mixture locally.

We obtain the following update relations:

$$\pi_k \quad \leftarrow \quad \frac{\sum_\theta p(\theta) \ P_K(k|\theta, \Gamma)}{\sum_\theta p(\theta)} \tag{18}$$

$$\mu_k \quad \leftarrow \quad \frac{\sum_\theta p(\theta) \ \theta \ P_K(k|\theta, \Gamma)}{\sum_\theta p(\theta) \ P_K(k|\theta, \Gamma)} \tag{19}$$

$$\sigma_k^2 \quad \leftarrow \quad \frac{\sum_\theta p(\theta) \ (\theta-\mu_k)^2 \ P_K(k|\theta, \Gamma)}{\sum_\theta p(\theta) \ P_K(k|\theta, \Gamma)} \tag{20}$$

where $P_K(\theta, k|\Gamma)$ is the posterior probability or the degree to which we suppose the data was generated by the Gaussian component k. One we have the data, it is computed with Bayes' rule:

$$P_K(k|\theta, \Gamma) = \frac{\pi_k \ \phi_k(\theta|\mu_k, \sigma_k)}{P_K(\theta|\Gamma)} \tag{21}$$

The accuracy of the EM may be influenced by the initial parameters, because of possible local maxima trap.

Our EM implementation re-used parts of Bouman's cluster package [9].

Our EM procedure operates as follows:

1. Initialization step
   • initialize K with a large number of classes K_{init}
   • initialize the weights equally, the means linearly, and the variances with the data variance:
   $K = K_{init}$, $\pi_k = 1/K$, $\sigma_k^2 = \text{var}(\theta)$ and
   $\mu_k = \theta(n)$ where $n = \lfloor (k-1)(N-1)/(K_{init}-1) \rfloor + 1$.
   • set a convergence threshold ϵ
2. Apply EM algorithm with Equations (21), (18), (19), (20)
   • compute $MDL(K, \Gamma)$ with Equation (17)
   • if change in $MDL(K, \Gamma)$ less than ϵ record $MDL(K, \Gamma)$, Γ
   • if $K > 1$, reduce K by merging the 2 nearest components, then $K \leftarrow K-1$ and go back to Apply EM step
   else stop (choose K^* and Γ^{K^*} than minimize the MDL value).
3. The number of sources estimate is K^*, and the located sources are the μ_k^* with priors π_k^* and variances σ_k^2.

4. SOURCE SEPARATION

4.1. Source Filtering Algorithm

In order to recover each source k, we select and regroup the time-frequency bins belonging to the same azimuth θ. We use two different masks to measure the proximity of each source.
The first allocates the energy of each bin to the source k according to its posterior probability (derived from EM). This is the probabilistic Maximum A-Posteriori (MAP) mask given with:

$$M_k(t, f) = P_K(k|\theta(t, f), \Gamma). \tag{22}$$

The second mask assigns the energy of each bin without sharing to the source k with the maximum Likelihood L_k. This is the binary mask based on maximum Likelihood given with:

$$M_k(t, f) = 1_{\{k=\text{argmax}_j \phi_j(\theta(t,f)|\theta_j, \sigma_j)\}} \text{ with } j = 1 \cdots K. \tag{23}$$

In practical terms, to avoid audible distortion, we consider a minimum mask value.

For each source k, the pair of short-term spectra are reconstructed following:

$$S_L(t,f) = M_k(t,f) \cdot X_L(t,f) \quad (24)$$
$$S_R(t,f) = M_k(t,f) \cdot X_R(t,f) \quad (25)$$

The time-domain version of each source k is obtained through a short-time inverse Fourier transform.

5. THE MOSPALOSEP APPLICATION

At the present time, the MOSPALOSEP platform is implemented under LabWindows/CVI with a graphical user interface (see Figure 3). In the *Input frame*, the platform allows the processing of a stereo signal samples coming from a wav file. The separated source signals can be broadcast through the sound card or saved in a wav file for future use (see *Output frame*).
In the *Spatialization and Mixing frame*, the stereo signal can also be synthesized artificially given the available spatialization methods (*e.g.* MHRIR using HRTFs or SSPA [10] using a parametric binaural model), which projects each source along the horizontal plane to a target azimuth. Listening of the previews is possible through the push-buttons provided for that purpose. Then we can mix the spatialized sources by dropping the *mix selected button*.
In the *localization frame*, we can choose a localization method among EM localizer, and different Histogram localizers, in the latter case, it is possible to level out and cap the power histogram. In the current implementation, we fix the resolution of the histogram to 361 discrete azimuths between $-180°$ and $180°$.
In the *Demixing frame*, the number of sources is necessary. We envisage a maximum of 4 sources in the mix. The EM parameters can be determined automatically or imposed by filling the edit text spaces. The probabilistic source separation with the MAP mask is the default setting, you can switch to the source separation using the ML mask. A performance measure is only possible with synthetic mixtures, since the original sources are essential for the computation of SNR and SIR. The Plot frame provides different views such as original signals in time or frequency domain, separated signals and histogram.

6. SIMULATION RESULTS

6.1. Localization Performance Analysis

We compare the localization and the number of sources estimation of the EM based localization (EM-loc) to the histogram based localization (Histo-loc), we make a difference between the localization with the raw histogram (Histo-loc-raw), with the leveled out histogram (Histo-loc-smooth) and with the leveled out and capped histogram (Histo-loc-threshold). The histograms are normalized to the maximum value, we use a threshold of -40 dB. The test signals are white noises spatialized to a target azimuth using SSPA.
We are exploring the case of one source mix. In our experiments, all signals are 2 seconds long with a sampling frequency of 44.1 kHz. We use a sliding Hanning window of 2048 samples and a 50% overlap between two consecutive windows. For the Histogram based methods, we choose a location (among the estimated locations) which is nearest to the theoretical location. In the EM-loc method, we choose the estimate with the highest prior probability.
The results show that the 3 Histo-loc approaches find the target azimuth with a average error of $1°$ (see Figure 4) over the azimuth space, while the EM-loc makes almost no error in the range $[-65°, +65°]$ (see Figure 4). Moreover, the EM-loc has a stable number of sources identification over all azimuths, it overestimated the number of sources by 2 to 3 sources. Despite the raw and leveled out Histo-loc methods overestimation of the number of sources, the Histo-loc-threshold is globally better, with a convergence on the exact number of sources when the source nears the center position ($0°$).

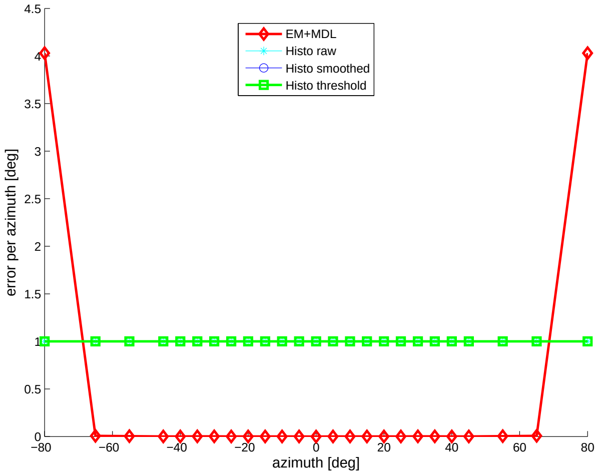


Figure 4: *Localization error per azimuth for different localization methods. Case of one-source mix.*

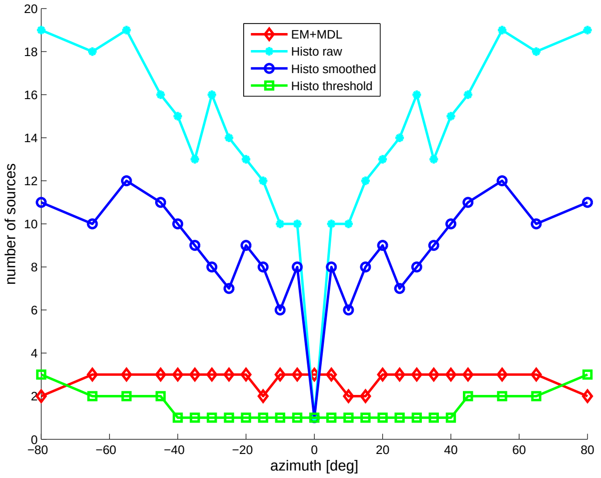
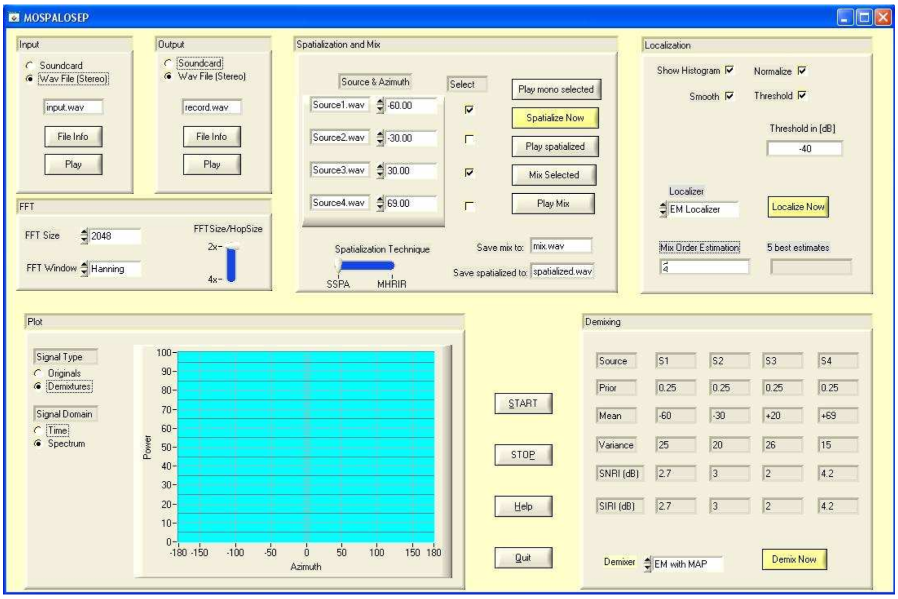


Figure 5: *Number of sources per azimuth for different number of sources identification methods. Case of one-source mix.*

6.2. Source Separation Performance Analysis

We evaluate the source separation algorithms using a probabilistic MAP mask and a ML mask. We used mix of 2 to 4 instrumental sources.



Input: ○ Soundcard ● Wav File (Stereo) — input.wav

Output: ○ Soundcard ● Wav File (Stereo) — record.wav

Source & Azimuth		Select
Source1.wav	-60.00	☑
Source2.wav	-30.00	☐
Source3.wav	30.00	☑
Source4.wav	69.00	☐

Spatialization Technique: SSPA / MHRIR

Save mix to: mix.wav

Save spatialized to: spatialized.wav

FFT Size: 2048

FFT Window: Hanning

FFTSize/HopSize: 2x– 4x–

Show Histogram ☑ Normalize ☑

Smooth ☑ Threshold ☑

Threshold in [dB]: -40

Localizer: EM Localizer

Mix Order Estimation: 4

5 best estimates:

Signal Type: ○ Originals ● Demixtures

Signal Domain: ○ Time ● Spectrum

Source	S1	S2	S3	S4
Prior	0.25	0.25	0.25	0.25
Mean	-60	-30	+20	+69
Variance	25	20	26	15
SNRI (dB)	2.7	3	2	4.2
SIRI (dB)	2.7	3	2	4.2

Demixer: EM with MAP

Figure 3: *The MOSPALOSEP Graphical User Interface under LabWindows/CVI.*

We use 4 monophonic sources (bass, percussion, guitar and piano) [11] with a sampling frequency of 44.1 kHz. The signals are spatialized artificially using Equations (5) and (6), and then mixed to form a binaural signal. Musical signals are particularly difficult to separate in comparison with speech signals, as they do not satisfy the WDO condition. Indeed, the spectrograms of the 4 sources show that the sources have a significant collision probability for frequencies in range $[0, 10]$ kHz (see Figure 6).

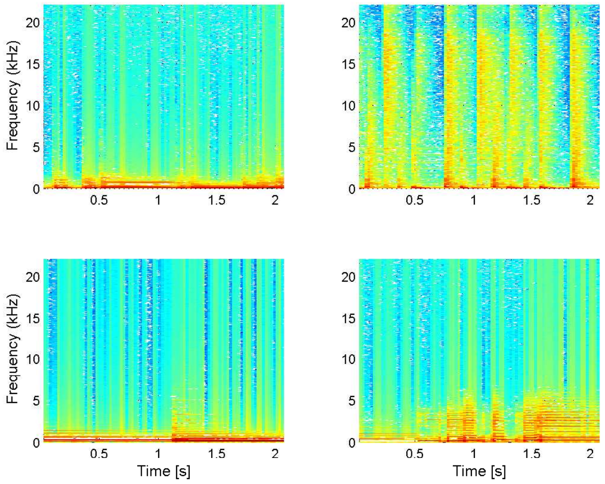


Figure 6: *Spectrograms of the four sources for simulations, from left to right, top to bottom: bass, percussion, guitar and piano .*

As in any source separation application, the ear is the first quality assessor of the extracted sources. In order to measure the robustness of the systems to interference and to global noise, we use the Signal to Interference Ratio (SIR) and the Signal to Noise Ratio (SNR), expressed in decibels (dB). In fact the SNR considers all discernable noises (musical noise, artifacts, interference). See [3] for details about SNR and SIR calculations. The SIRI and SNRI represent the SIR and SNR Improvement between the input and output SIR and SNR values.

In the case of two sources, there are three possible source combinations s_1, s_2 or (s_1, s_2). A maximum of two sources contribute to the energy of the mix point (t, f). The binary ML mask assigns the frequency energy to s_1 or s_2 without taking into account the case where both sources are active. While the MAP mask always anticipates an overlap situation. In the ideal case, the MAP mask is analogous to the ML mask. For the source separation stage, we use a modified EM approach; in fact for the GMM computation, only the set of discrete azimuths covered by the leveled out histogram are used with their corresponding average energy.

The Figure 7 depicts a typical result of separation. The *bass* at $0°$ and the *percussion* at $15°$. We note that the MAP mask based estimates present less interference, but have energy levels lower than the originals, which could lead to some distortion. Estimates from the ML mask have less distortion, but they are marred by much interferences. Figure 7 confirms that the MAP mask is superior to ML mask for an equivalent input level of distortion. The average SIR gains are higher than 12 dB, with a SNR gain of approximately 7 dB (see Table 1).

Moreover, informal listening tests reveal that MAP masks separated sources are preferable to those from the ML mask[1].

In the case of 3 and 4 sources, we have respectively 7 and 15 possible source combinations, thus a higher collision probability. Table 1 certifies that SIR levels increase with the source number, and the quality of estimates is also affected. The SIR gain remains higher than 10 dB, but the SNR decreases slightly compared to the case of two-source mix. The SIR improvement, despite of more sources, can be explained by the fact that the input SIR is too low,

[1] http://recherche.esigetel.fr/~akg

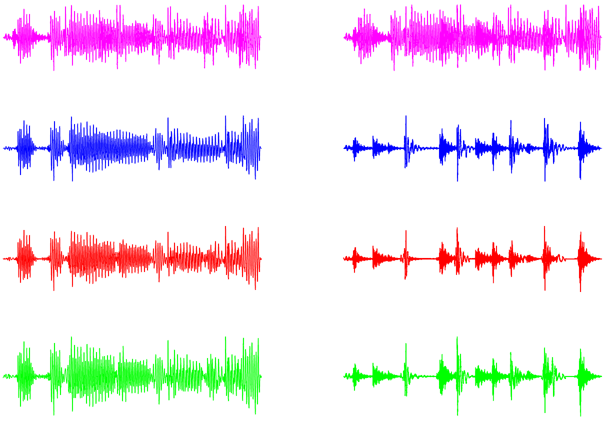


Figure 7: *(row 1) Mixtures of two instrumental sources (bass, 0°) and (percussion, -15°), (row 2) originals, (row 3) separated sources with MAP mask, (row 4) separated sources with ML mask.*

source (θ)	SIR in (dB)	SIRI (dB)	SNRI (dB)
$bass\ (0^\circ)$	8.34	9.36	1.02
$percussion\ (+45^\circ)$	-7.19	18.71	8.77
$bass\ (-20^\circ)$	-6.89	11.07	12.73
$percussion\ (0^\circ)$	-2.93	19.48	12.03
$piano\ (+15^\circ)$	-0.69	16.53	8.07
$bass\ (-15^\circ)$	-3.36	16.41	8.6
$percussion\ (0^\circ)$	-2.15	12.34	7.4
$guitar\ (15^\circ)$	-3.07	14.05	7.75
$piano\ (+30^\circ)$	-11.35	15.26	12.07

Table 1: *Source separation performance for the MAP mask given binaural mixtures with 2, 3 or 4 sources. We measured the SIR and the SNR, and the obtained gains. We have SIR gains greater than 10 dB, and SNR gains greater than 5 dB.*

and a improvement provides a real gain in value. The tests indicate that the ML mask becomes less useful with a rising number of sources. For example, the superposition of two opposing side sources can give rise to a point between the two sources, which corresponds to the third source, and makes the separation more complex.

We also explore four-source mix, and we notice the same observations as in the three-source case. Still, we have interference gain greater than 10 dB and a slight degradation of SNR. The separation performance seems to be nearing its limit (Table 1). The overall quality of estimates has also deteriorated audibly. These tests demonstrate that the probabilistic posterior mask overcomes the binary mask in complex signal separation.

7. CONCLUSIONS AND FUTURE WORK

We presented the MOSPALOSEP platform for binaural source localization and separation. We showed comparison results between EM based localization and histogram based localization. We made an analysis of the number of sources identification, and showed that the EM with MDL is a good candidate for this purpose in audio applications. We also compared different source separation methods, one using a posterior probability mask that takes into account the possibility of a superposition amongst sources. This latter overcomes the binary mask separation. The results showed significant gains in terms of interference and distortion reduction. Next, we will enrich the toolbox with other methods and performance metrics referenced in the literature for more comparisons. We hope to study the influence of the spatial resolution of sound sources, and also to investigate the localization and separation methods in case of added spatial-distributed noise on both signals. We plan also to give more insight in the cases where the WDO constraint is severely violated.

8. ACKNOWLEDGMENTS

Many thanks to Sylvain Marchand, university of Bordeaux.

9. REFERENCES

[1] H. Viste, *Binaural Localization and Separation Techniques*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[2] C. Avendano, “Frequency-Domain Source Identification and Manipulation in Stereo Mixes for Enhancement, Suppression, and Re-Panning Applications,” in *Proc. IEEE WASPAA*, New York, 2003.

[3] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Trans. on Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[4] J. Mouba and S. Marchand, “A source localization/separation/respatialization system based on unsupervised classification of interaural cues,” in *Proc. Digital Audio Effects (DAFx-06)*, Montréal, Canada, Sept. 18- 20, 2006, pp. 233–238.

[5] K. Hirokazu, N. Takuya, and S. Shigeki, “Separation of Harmonic Structures Based on Tied Gaussian Mixture Model and Information Criterion for Concurrent Sounds,” in *Proc. IEEE ICASSP*, Montreal, 2004, pp. 297–300.

[6] J. Blauert, *Spatial Hearing*, MIT Press, Cambridge, Massachusetts, revised edition, 1997, Translation by J. S. Allen.

[7] J. Mouba, S. Marchand, B. Masencal, and J-M. Rivet, “Retrospat: a perception-based system for semi-automatic diffusion of acousmatic music,” in *Proceedings of the Sound and Music Computing (SMC)*, Berlin, Germany, Sept. 31- Aug. 3, 2008, pp. 33–40.

[8] J. Rissanen, “A Universal Prior for Integers and Estimation by Minimum Description Length,” *Annals of Statistics*, vol. 11, no. 2, pp. 417–431, 1983.

[9] C. A. Bouman, “Cluster: an unsupervised algorithm for modeling gaussian mixtures,” Tech. Rep., Purdue University, 1995.

[10] J. Mouba, “Performance of source spatialization and source localization algorithms using conjoint models of interaural level and time cues,” in *Proc. Digital Audio Effects (DAFx-09)*, Como, Italy, Sept. 1- 4, 2009.

[11] E. Vincent, R. Gribonval, and C. Févotte, “http://www.irisa.fr/metiss/bass-db,” Tech. Rep., Académie Montpellier, 2005.