A SOURCE LOCALIZATION/SEPARATION/RESPATIALIZATION SYSTEM BASED ON UNSUPERVISED CLASSIFICATION OF INTERAURAL CUES

Joan Mouba, Sylvain Marchand

SCRIME – LaBRI, University of Bordeaux 1
351 cours de la Libération, F-33405 Talence cedex, France
firstname.name@labri.fr

ABSTRACT

In this paper we propose a complete computational system for Auditory Scene Analysis. This time-frequency system localizes, separates, and spatializes an arbitrary number of audio sources given only binaural signals. The localization is based on recent research frameworks, where interaural level and time differences are combined to derive a confident direction of arrival (azimuth) at each frequency bin. Here, the power-weighted histogram constructed in the azimuth space is modeled as a Gaussian Mixture Model, whose parameter structure is revealed through a weighted Expectation Maximization. Afterwards, a bank of Gaussian spatial filters is configured automatically to extract the sources with significant energy accordingly to a posterior probability. In this frequency-domain framework, we also inverse a geometrical and physical head model to derive an algorithm that simulates a source as originating from any azimuth angle.

1. INTRODUCTION

In many applications, it may be desirable to manipulate the individual sound sources that can be heard in a mix. It remains a great challenge to separate the sources in our case where we only have two sensors (human ears) and apply no restriction on the number of sources. In this case, called *degenerated*, most of the techniques for source separation, based on matrix inversion, just fail.

Another approach tries to mimic the human auditory system, based on perception and psychoacoustics: Computational Auditory Scene Analysis techniques (CASA). A recent source separation approach called *Degenerate Unmixing and Estimation Technique* (DUET) was proposed by Rickard *et al.* [1]. This technique relies on binaural cues – the interaural differences in time (ITD) and amplitude (ILD) – which play an important role in the human localization system, since they are both related to the azimuth of the source. However, DUET is mainly restricted to low frequencies, since the phase becomes ambiguous above 1500 Hz. Viste and Evangelista [2] get rid of this ambiguity by minimizing the distance between the ILD and ITD based azimuth estimates, thus obtaining an enhanced azimuth estimate for each time-frequency bin. The power of each bin is then accumulated in an histogram, where we can read the energy as a function of the azimuth.

Furthermore, in order to separate sources, Rickard *et al.* attribute exclusively the energy of a bin to one source. In contrast, Avendano proposes in [3] an adaptive spatial Gaussian mapping to achieve the source separation. However, in his approach there is no true azimuth but only an inter-channel amplitude panning coefficient for each source. Moreover, the inter-channel time / phase is not considered.

Here we introduce a Gaussian Mixture Model (GMM) of the azimuthal histogram, and use a Maximum Likelihood (ML) approach based on a modified Expectation Maximization (EM) [4] to learn the GMM structure (mix order, weight, azimuthal location and deviation of each source) from the power-weighted histogram. A similar reformulation of EM was proposed in [5] in the case of spectral density functions.

The GMM parameters setup automatically a demixing stage (Figure 1) where bins belonging to each source are statistically selected and the energy of each bin is assigned according to a posterior probability. In contrast to others, we consider the exclusive energy assignment as too destructive.

This paper is organized as follows. In Section 2, we describe the binaural model, the localization approach, and we also detail the time-frequency algorithm proposed to map a source signal to a pair of binaural signals with the expected Direction Of Arrival (DOA). Section 3 introduces a GMM of the energy-weighted histogram and explains the parameters learning with an EM approach. In the same section, we present the probabilistic demixing algorithm. Finally, experiments and results are described in Section 4.

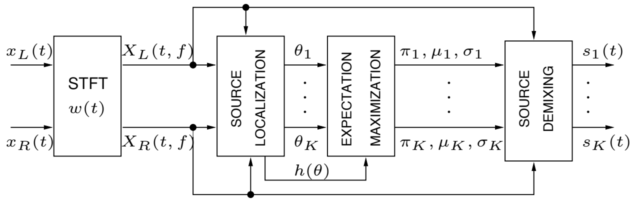

Figure 1: *Overview of the proposed CASA-EM system.*

2. SINGLE SOURCE

2.1. Model

A (vibrating) sound source radiates spherical acoustic waves, that propagate to the ears through an energy transmission between air particles of the surrounding environment. In this paper, we consider the sound sources as punctual and omni-directional.

In a polar coordinate system (see Figure 2), the source point is localized given its (ρ, θ, ϕ) coordinates, where ρ is the distance between the source and the head center (O), θ is the azimuth angle, and ϕ the elevation angle.

In the present work, the sources are approximately in the same horizontal plane as the ears ($\phi = 0$). This is the case in many musical situations, where both the listener and instrumentalists are standing on the (same) ground. Also, we focus on the DOA and thus neglect the distance between the source and the listener. More precisely, we are in a situation where the instrumentalists are a few meters from the listener. We consider that this distance is small enough to neglect the (frequency-selective) attenuation by the air, though large enough for the acoustic wave to be regarded as planar when reaching the ears. As a consequence, our source localization

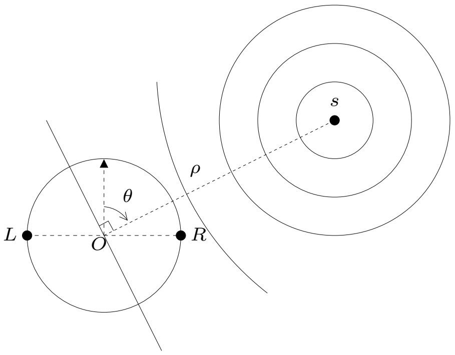

Figure 2: *A source s positioned in the horizontal plane at azimuth θ, propagating acoustic waves to the head.*

/ separation / respatialization system will depend on the azimuth angle θ only.

Moreover, we consider that we are in outdoors condition. We consider neither any room nor any obstacle. In this free-field case, only the head, the torso, and the outer-ear geometry modify the sound activity content by reflections and shadowing effect.

The source s will reach the left (L) and right (R) ears through different acoustic paths, characterizable with a pair of Head-Related Impulse Responses (HRIR). For a source s located at the azimuth θ, the left (x_L) and right (x_R) signals are given by:

$$x_L = s * \mathrm{HRIR}_L(\theta) \quad (1)$$

$$x_R = s * \mathrm{HRIR}_R(\theta) \quad (2)$$

where $*$ is the convolution among time-domain signals, the HRIR being impulse responses of filters. The HRIRs depend on the morphology of the head of the listener, are different for each ear, and are functions of the location of the source as well as its frequency.

In fact, each HRIR can be regarded as a filter, and defined in the time domain by its impulse response for each azimuth θ. The CIPIC database [6] contains this information for several listeners and different directions of arrival. In our experiments, to be listener-independent, we consider the mean HRIR (MHRIR), that is, for a given position θ, the average (in the time domain) of the HRIRs for this position among the several listeners.

In the case of s being a pure sinusoid, the convolution in preceding equations is replaced by a simple multiplication in the spectral domain. A sound source positioned to the left will reach the left ear sooner than the right one, in the same manner the left level should be higher due to head-shadowing.

More precisely, between the two ears, the amplitude ratio is approximately 10^{Δ_a} and the phase difference is approximatively Δ_ϕ, where Δ_a and Δ_ϕ are given by:

$$\Delta_a = \mathrm{ILD}(\theta, f)/20 \quad (3)$$

$$\Delta_\phi = \mathrm{ITD}(\theta, f) \cdot 2\pi f \quad (4)$$

where the difference in amplitude or interaural level difference (ILD, expressed in decibels – dB) and difference in arrival time or interaural time difference (ITD, expressed in seconds) are the principal spatial cues for the human auditory system localization. Since we restrict our preliminary study to the simple free-field case, the inter-channel coherence [7] will not be considered.

These binaural cues can be related to physical parameters such as the celerity of sound c and the head radius r. From the analysis of the CIPIC database, Viste [2] extends the Woodworth's formula [8] for the ITD and derives a sinusoidal model for the ILD:

$$\mathrm{ILD}(\theta, f) = \alpha_f \sin(\theta) \quad (5)$$

$$\mathrm{ITD}(\theta, f) = \beta_f r \left(\sin(\theta) + \theta\right)/c \quad (6)$$

where α_f and β_f are frequency-dependent scaling factors (see [2]), that encapsulate the head / ears morphology. In our experiments, we use the mean of individual scaling factors over the 45 subjects of the CIPIC database. For each subject, we measure the interaural cues from the HRIR and derive the individual scaling factors that best match the model – in the least-square sense – for all azimuths.

In the case of a complex sound, we can decompose it as a sum of sinusoids. We perform a time-frequency decomposition using the Short-Time Fourier Transform (STFT):

$$X(t, f) = \int_{-\infty}^{\infty} x(t+\tau) w(\tau) e^{-j2\pi f\tau} d\tau \quad (7)$$

where x is the temporal signal, w some windowing function, and X is the (short-term) spectrum of x.

In our practical experiments, we use sounds sampled at 44100 Hz. We then perform discrete Fourier transforms of size $N = 2048$ using the Fast Fourier Transform (FFT) algorithm. For w, we use the periodic Hann window. The hop size between two consecutive windows is $H = N/2$ (50% overlap). For each spectrum, we use the zero phase technique: In order to cancel the linear phase of the analysis window, we swap the two halves of the sample buffer prior to any FFT (and after any Inverse FFT).

Then each point of the time-frequency plane can be regarded as the contribution of a single sinusoid, and the Δ_a and Δ_ϕ coefficients can in turn be used, as functions of frequency. In fact, we should have the following relation between the spectra measured at the left and right ears:

$$X_R(t, f) = X_L(t, f) \cdot 10^{\Delta_a} e^{j\Delta_\phi} \quad (8)$$

2.2. Localization

In Auditory Scene Analysis, the ILD and the ITD are the most important cues for source localization. Lord Rayleigh (1907) mentioned in his Duplex Theory that the ILD are more prominent at high frequencies whereas the ITD are crucial at low frequencies. In fact, the human auditory system is well adapted to the natural environment. Indeed, high frequencies are more sensitive to frequency-selective amplitude attenuation (by the air or the head shadowing), but the associated signals exhibit phase ambiguities. In contrast, low frequencies are not ambiguous, but are less sensitive to amplitude attenuation.

Given the short-time spectra of the left and right channels, we can estimate the ILD and ITD for each time-frequency bin with:

$$\mathrm{ILD}(t, f) = 20\log_{10}\left|\frac{X_R(t, f)}{X_L(t, f)}\right| \quad (9)$$

$$\mathrm{ITD}_p(t, f) = \frac{1}{2\pi f}\left(\angle\frac{X_R(t, f)}{X_L(t, f)} + 2\pi p\right) \quad (10)$$

The coefficient p outlooks that the phase is determined up to a modulo 2π factor. In fact, the phase becomes ambiguous beyond 1500 Hz, according to the Duplex Theory.

Obtaining an estimation of the azimuth based on the ILD information is just a matter of inverting Equation (5):

$$\theta_L(t, f) = \arcsin\left(\frac{\mathrm{ILD}(t, f)}{\alpha_f}\right) \quad (11)$$

Similarly, using the ITD information, to obtain an estimation of the azimuth candidate for each p, we invert Equation (6) by computing:

$$\theta_{T,p}(t,f) = \Pi\left(\frac{c \cdot \text{ITD}_p(t,f)}{r \cdot \beta_f}\right) \quad \text{with}$$

$$\Pi(x) = 0.50018\,x + 0.009897\,x^3 + 0.00093\,x^5 + O(x^5) \quad (12)$$

$\Pi(x)$ is an order-5 polynomial approximation of the inverse of the $\sin(\theta) + \theta$ function. The $\theta_L(t,f)$ estimates are more dispersed, but not ambiguous at any frequency, so they are exploited to find the right modulo coefficient p that unwraps the phase. Then the $\theta_{T,p}(t,f)$ that is nearest to $\theta_L(t,f)$ is validated as the final θ estimation, since it exhibits a smaller deviation:

$$\theta(t,f) = \theta_{T,m}(t,f) \text{ with } m = \text{argmin}_p\,|\theta_L(t,f) - \theta_{T,p}(t,f)| \quad (13)$$

In theory, in the case of a single source all frequencies should give the same azimuth, exactly corresponding to the source position θ. However, in practice, the presence of noise and estimation errors make things a little more complicated. In fact, as a first approximation, we will consider that the energy of the source is spread following a Gaussian distribution centered at the theoretical value θ. Although the Gaussian nature of the distribution is questionable and should be verified in the near future, we are comforted by the well-known Central Limit Theorem as well as preliminary experiments. In this context, the ideal case is a Gaussian of mean θ and variance 0.

2.3. Spatialization

In order to spatialize a sound source s to an expected position θ (see Figure 3), we first transform its (mono) signal into the time-frequency domain using a windowed FFT. Then the pair of left and right spectra is computed. A first approximation is to set $X_L = S$ (the spectrum of s), and to compute X_R from X_L using Equation (8). This trivially respects the interaural cues, but one of the ears plays a specific role then. Another solution is to divide the spatial cues equally, using the Equations (3), (4), (5), (6), and the following equations:

$$X_L(t,f) = X(t,f) \cdot 10^{-\Delta_a/2} e^{-j\Delta_\phi/2} \quad (14)$$

$$X_R(t,f) = X(t,f) \cdot 10^{+\Delta_a/2} e^{+j\Delta_\phi/2} \quad (15)$$

where X is the spectrum reaching both ears when the source is played from position $\theta = 0$. More precisely, $X = S \cdot \text{HRTF}(0)$, where HRTF – Head-Related Transfer Function – is the spectral equivalent of the HRIR (S, the spectrum of the source itself, being directly perceived by the left or right ear if θ is $-\pi/2$ or $\pi/2$, respectively). Nevertheless, as a first approximation, we will take $X = S$, thus the interaural cues will be respected, as well as the symmetric role of the ears.

Finally, the signals x_L and x_R are computed from their spectra X_L and X_R using the Inverse FFT, and sent to the left and right ears via headphones. In practice, we use the periodic Hann window with an overlap factor of 25%.

3. SOURCE SEPARATION

3.1. WDO Assumption

To achieve degenerated separation of a arbitrary number of sources given binaural mixtures, we consider any pair of sources $(s_k(t), s_l(t))$ as Windowed-Disjoint Orthogonal (WDO). This means that their short-time spectra do not superpose. This ideal case is given by:

$$\forall k \neq l, \quad S_k(t,f) \cdot S_l(t,f) = 0 \qquad (k,l = 1,\cdots,K) \quad (16)$$

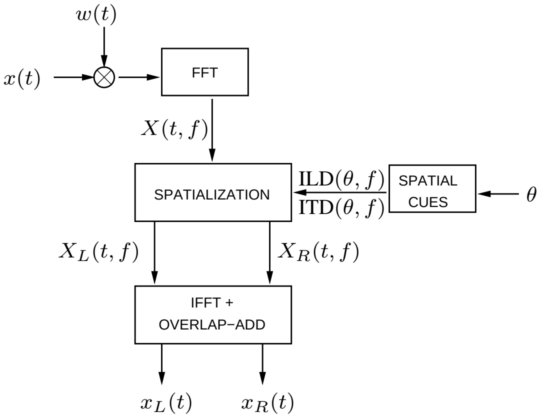

Figure 3: *Spatialization of a source x to an azimuth θ.*

where K is the number of sources in the mix. This condition is rarely satisfied in music signals, at least exactly. However, experiments carried out in [1] verify that speech signals are approximatively WDO.

3.2. Histogram Building

We define a discrete axis for the azimuth values with resolution Δ_θ. Every computed $\theta(t,f)$ is compared to its nearest discrete neighbor multiple of Δ_θ, in order to define a binary mask:

$$M_\theta(t,f) = \begin{cases} 1 & \text{if } |\theta(t,f) - \theta| \leq \Delta_\theta/2 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

At each time t, we then cumulate the power of the source in an histogram:

$$h(\theta) = \sum_f |M_\theta(t,f) X_L(t,f) X_R(t,f)| \quad (18)$$

Note that, from Equations (14) and (15), the energy we cumulate is $X_L X_R = X^2$, which would be roughly the energy of the source if it were played at $\theta = 0$ (monophonic case where $x_L = x_R$). This makes the histogram unbiased: Rotating the source is equivalent to a shift of the θ axis of the histogram, but then its energy remains unchanged.

3.3. Gaussian Mixture Model

Since the sources are not exactly WDO, for each source we obtain a distribution around the true value. As mentioned in Section 2, we choose to approximate the energy accumulation distribution around each point with a Gaussian distribution. In the case of K sources, we then introduce a model of K Gaussians (K-GMM, order-K Gaussian mixture model):

$$P_K(\theta|\Gamma) = \sum_{k=1}^{K} \pi_k\,\phi_k(\theta|\mu_k,\sigma_k) \text{ with } \pi_k \geq 0 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \quad (19)$$

where Γ is a multiset of K triples $(\pi_k, \mu_k, \sigma_k^2)$ that 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 component described mathematically by:

$$\phi_k(\theta|\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(\theta-\mu_k)^2}{2\sigma_k^2}\right) \quad (20)$$

We are interested in estimating the architecture of the K-GMM, that is the number of sources K and the set of parameters Γ, to be able to setup the separation filtering.

3.3.1. First Estimation

In the histogram, we observe local maxima which number provides an estimation of the number of sources in the mixture. The abscissa of the k-th local maximum reveals the location θ_k of the k-th source. Generally, a finer tuning of the histogram allows a visual separability of very close-lying sources. But we will detect other local maxima around the true maxima. So, we propose to smooth the histogram using a binomial operator $\mathcal{B}$ to fuse the encountered peaks:

$$\mathcal{B}(n) = \frac{1}{2^{D-1}\binom{D-1}{n}} \qquad n = 0, \cdots, D-1 \tag{21}$$

where D is the dimension of the operator. We recommend to use an odd order bigger than three. The smoothed histogram $\tilde{h}$ is constructed by convolution of the current histogram h with the smoothing kernel $\mathcal{B}$.

The interferences between sources bring corrupted cues, from which the cumulated energy indicates a source at a position that does not match that of any existing source. When the sources do not coincide, usually the heights of the undesired peaks are very low. Thus, we apply a threshold, such that only the peaks above are considered as real sources. The floor of the threshold is set relatively to a noise level estimate in the histogram.

Informal experiments show that the estimated source number and location are rather good. This gives the model order K and a first estimation of the means of the Gaussians (μ_k in Γ). This estimation can be refined and completed – with the variances σ_k^2 and the weights π_k – for example by the EM algorithm.

3.3.2. Expectation Maximization

Each source in the mix is characterized with a Gaussian representation in the histogram. To discriminate the sources, the weight, mean, and variance of each source are essential for our spatial Gaussian filtering algorithm.

Expectation Maximization (EM) is a popular approach to estimate parameters in mixture densities given a data set x. The idea is to complete the observed data x with a unobserved variable y to form the complete data (x, y), where y indicates the index of the Gaussian component from which x has been drawn. Here, the role of x is played by the azimuth θ, taking values in the set of all discrete azimuths covered by the histogram. We associate θ with its intensity function $\tilde{h}(\theta)$ (the smoothed histogram). The role of y is played by $k \in \{1, \cdots, K\}$, the index of the Gaussian component θ should belong to.

The EM algorithm proceeds iteratively, at each iteration the optimal parameters that increase locally the log-likelihood of the mixture are computed. In other words, we increase the difference in log-likelihood between the current with parameters Γ and the next with parameters Γ'. This log-form difference, noted $Q(\Gamma', \Gamma)$, can be expressed as:

$$\begin{aligned} Q(\Gamma', \Gamma) &= \sum_\theta \tilde{h}(\theta)\left(\mathcal{L}(\theta|\Gamma') - \mathcal{L}(\theta|\Gamma)\right) \quad \text{with} \\ \mathcal{L}(\theta|\Gamma) &= \log\left(P_K(\theta|\Gamma)\right) \end{aligned} \tag{22}$$

We can then reformulate $\mathcal{L}(\theta|\Gamma)$ like this:

$$\begin{aligned} \mathcal{L}(\theta|\Gamma) &= \log\left(\sum_k P_K(\theta, k|\Gamma)\right) \quad \text{with} \\ P_K(\theta, k|\Gamma) &= \pi_k\ \phi_k(\theta|\mu_k, \sigma_k) \end{aligned} \tag{23}$$

The concavity of the log function allows to lower bound the $Q(\Gamma', \Gamma)$ function using the Jensen's inequality. We can then write:

$$Q(\Gamma', \Gamma) \geq \sum_\theta \sum_k \tilde{h}(\theta) P_K(k|\theta, \Gamma) \log\left(\frac{P_K(\theta, k|\Gamma')}{P_K(\theta, k|\Gamma)}\right) \tag{24}$$

where $P_K(k|\theta, \Gamma)$ is the posterior probability, the degree to which we trust that the data was generated by the Gaussian component k given the data; it is estimable with the Bayes rule:

$$P_K(k|\theta, \Gamma) = \frac{P_K(\theta, k|\Gamma)}{P_K(\theta|\Gamma)} \tag{25}$$

The new parameters are then estimated by maximizing the lower bound with respect to Γ:

$$\Gamma' = \text{argmax}_\gamma \sum_\theta \sum_k \tilde{h}(\theta) P_K(k|\theta, \Gamma) \log\left(P_K(\theta, k|\gamma)\right) \tag{26}$$

Increasing this lower bound results automatically in an increase of the log-likelihood, and is mathematically easier. Finally, the maximization of Equation (26) provides the following update relations (to be applied in sequence, because they modify – update – the current value with side-effects, thus the updated value must be considered in the subsequent relations):

$$\pi_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta, \Gamma)}{\sum_\theta \tilde{h}(\theta)} \tag{27}$$

$$\mu_k \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ \theta\ P_K(k|\theta, \Gamma)}{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta, \Gamma)} \tag{28}$$

$$\sigma_k^2 \leftarrow \frac{\sum_\theta \tilde{h}(\theta)\ (\theta - \mu_k)^2\ P_K(k|\theta, \Gamma)}{\sum_\theta \tilde{h}(\theta)\ P_K(k|\theta, \Gamma)} \tag{29}$$

The performance of the EM depends of the initial parameters. The first estimation parameter should help to get around likelihood local maxima trap. Our EM procedure operates as follows:

1. Initialization step
   - initialize K with the order of the first estimation
   - initialize the weights equally, the means according to the first estimation, and the variances with the data variance (for the initial Gaussians to cover the whole set of data): $\pi_k = 1/K, \quad \mu_k = \theta_k, \text{ and } \sigma_k^2 = \text{var}(\theta)$
   - set a convergence threshold ϵ
2. Expectation step
   - compute $P_K(k|\theta, \Gamma)$ with Equation (25)
3. Maximization step
   - compute Γ' from Γ with Equations (27), (28), and (29)
   - if $P_K(\theta|\Gamma') - P_K(\theta|\Gamma) > \epsilon$
   then $\Gamma \leftarrow \Gamma'$ and go back to the Expectation step
   else stop (the EM algorithm has converged).

3.4. 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 the parameters issued from the EM-component number k, and the energy of the mixture channels is allocated to the (left and right) source channels according to the posterior probability. More precisely, we define the following mask for each source:

$$M_k(t,f) = P_K(k|\theta(t,f),\Gamma) \tag{30}$$

if $10\log_{10}|\phi_k(\theta(t,f)|\mu_k,\sigma_k)| > L_{\text{dB}}$, and 0 otherwise. This mask limits the fact that the tail of a Gaussian distribution stretchs out to infinity. Below the threshold L_{dB} (expressed in dB, and set to -20 in our experiments), we assume that a source of interest does not contribute anymore. For each source k, the pair of short-term spectra can be reconstructed according to:

$$\begin{aligned} S_L(t,f) &= M_k(t,f)\cdot X_L(t,f) && (31)\\ S_R(t,f) &= M_k(t,f)\cdot X_R(t,f) && (32)\end{aligned}$$

The time-domain version of each source k is finally obtained through an inverse Fourier transform and an overlap-add procedure (as described in Section 2).

4. SIMULATION RESULTS

We have implemented, using MATLAB, two source separation / localization systems: DUET and the method proposed in Section (3), as well as two source spatialization techniques: MHRIR based on Equations (1) and (2) and the method we proposed in Section 2, called SSPA (see Figure 3).

4.1. Spatialization Results

To verify the effectiveness of the SSPA and MHRIR spatialization systems, we spatialized speech and music sources, then we conducted listening tests. Since determining the absolute location of a source could be a hard task for a listener, we preferred to conduct tests on relative locations. One (mono) source was spatialized to two different locations to create two (stereo) sounds – each sound consisting of a pair of binaural signals. During the audition, listeners had to tell if the second sound was at the left, right, or same position in comparison to the first sound. In a first round, we considered only couples of sounds produced with the SSPA system (parameterized with a sliding Hann window of length 2048 samples and a 25% overlap). The analysis of the results indicates no ambiguity between left and right. However, sources distant from less than 5° were often judged as coming from the same location. For comparison purposes, we also attended cross hearing experiments. The disposition was similar to the previous one, but here one of the source was spatialized using the MHRIR method. First observation, the MHRIR sources clang more naturally. The SSPA signals have more treble than the MHRIR signals. It is also likely that the SSPA still misses some characteristics of the perception. Another point, the MHRIR lateralized the source more accurately to the expected location. For example, a speaker voice spatialized to $-80°$ with the MHRIR was perceived more to the left than with the SSPA procedure. However, the HRIRs are known for only some discrete position (25 azimuths in the case of the CIPIC database), and interpolating them for the other positions is not trivial. SSPA does not suffer from this limitation. Enhancing the quality of SSPA is one of our research directions.

source	theory	1^{st} est.	EM-est.
xylophone	$-55°$	$-42°$	$-48°$
horn	$-20°$	$-13°$	$-15°$
kazoo	$30°$	$29°$	$25°$
electric guitar	$65°$	$64°$	$61°$

Table 1: *Locations of four sources estimated with the local maxima search, and with Expectation Maximization.*

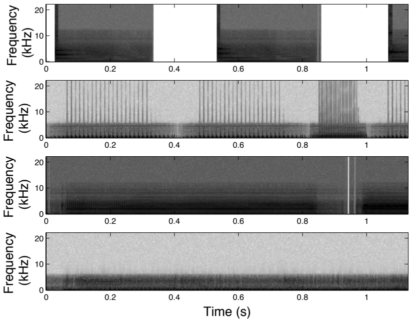

Figure 4: *Spectrograms of the sources of Table 1, from top to bottom: xylophone, horn, kazoo, and electric guitar.*

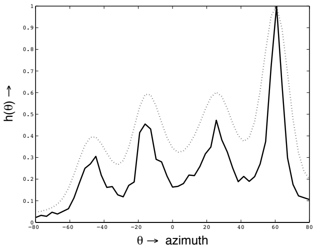

Figure 5: *Localization of four sources: histogram (solid line) and smoothed histogram (dashed line).*

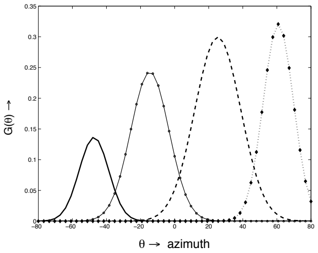

Figure 6: *GMM for the histogram of the four-source mix.*

4.2. Localization Results

Source localization is a decisive task for source separation based on spatial filtering. First, we created binaural signals with MHRIR, then the individually-created signals are mixed into a single binaural signal. In Figure 6, we show an example for the localization of four instruments in a binaural mixture with the proposed localization method: xylophone at $-55°$, horn at $-20°$, kazoo at $30°$, and electric guitar at $65°$. We note that the procedure performs well for music signals. However, we can hardly compare these results with the output of the DUET method, which relies on the interaural cues but ignore the azimuth. In the smoothed histogram version of the relative noisy histogram (Figure 5), many unwelcome peaks have been dropped, while the correct sources were enhanced. In the two-dimensional histogram (ILD and ITD) obtained with the DUET method, we observed a stockpile of peaks with substantial energy. More precisely, DUET is unable to handle the phase ambiguity (appearing above 1500 Hz) for our test signals sampled at 44100 Hz, with broadband content (Figure 4).

We then proceed to the localization using a direct local maxima search in the smoothed histogram (see Table (1)). It appeared that a size of 65 was sufficient for the histogram. To build this histogram, we used FFTs of 2048 samples with an overlap of 50%. The mix order was accurately identified (4 sources). The kazoo and the electric guitar were precisely localized, only an error of 1° is observable. As a matter of fact, the kazoo has constantly high power and a wide-band spectrum, so its peak is predictably high. But because their spectrograms overlap (see Figure (4)), the sources underly interferences that corrupt the cues. However, we still get an histogram that conveys an outgoing material to segregate the sources. In a next step, we modeled the histogram as a GMM. The normalized estimated Gaussian components are depicted in Figure 6. The individual peaks are incorporated inside a Gaussian distribution. This emphasizes our motivation that a Gaussian mixture is objective for separation purposes. The location estimates of EM are exposed in Table (1), and do not really enhance the first estimation of the Gaussian means. However, the EM algorithm gives us the widths of the Gaussian distributions. Moreover, in the near future we plan to enhance these results by applying the EM algorithm on the raw azimuth estimates $\theta(t, f)$ instead of the data stored in the smoothed histogram. Indeed, the operations consisting in setting the histogram size, building this histogram, then smoothing it, are well-suited for a direct local maxima search in the histogram, but these operations may alter the data themselves.

4.3. Separation Results

We carried out subjective investigations on the time-domain signals and spectrograms. And we also judged the similarity of the sources based on hearing test.

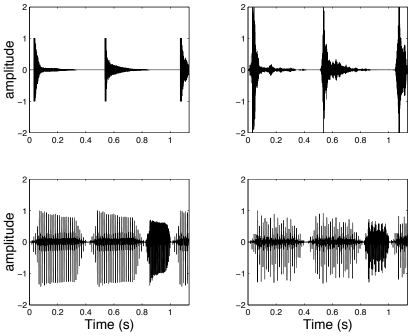

Figure 7: *Waveforms of the demixtures (on the right, originals being on the left): xylophone (*$-55°$*) (top) and horn (*$30°$*) (bottom).*

To separate the sources, a spatial filtering identifies and clusters bins attached to the same source. Many methods, like DUET, separate the signals by assigning each of the time-frequency bins to one of the sources exclusively. We assume that several sources can share the power of a bin, and we attribute the energy according to a membership ratio – a posterior probability. The histogram learning with EM provides a set of parameters for each Gaussian distribution. The parameters are then used to parameterize automatically a set of spatial Gaussian filters. A result of demixing is depicted in Figure (7) for a two-instrument mixture: xylophone at $-55°$ and horn at $-20°$; their original spectrograms are shown in Figure 4. In the time domain, the xylophone rhythm is respected, its signal looks amplified and its shape is preserved. Perceptively, the demixed xylophone is very similar to the original one. Also, for the horn, we must tolerate some interference effects, and the spectrograms are partly damaged. A portion of energy was absorbed by an unwanted source generated from interferences. We also conducted tests on speech samples. The reconstruction quality was good. The quality was much better than for long-distance telephone lines. Several listening examples for the spatialization and the separation systems are available online[1].

5. CONCLUSIONS AND FUTURE WORK

The performance tests show that our system reaches promising results for speech and music signals. In our future works we will extend the localization to moving sources, but we await undesirable phase effects in the re-synthesis. In the case of FFT bins shared by several sources, the source separation is an interesting challenge. We also plan to study the brightness of spectra to judge the closeness of a source. Currently, the overall system is being implemented in a real-time environment for many applications and further investigations as part of the InSpect/ReSpect projects at the SCRIME/LaBRI research centers in Bordeaux.

6. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] H. Viste, "Binaural localization and separation techniques," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[3] C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression, and re-panning applications," in *Proc. IEEE Workshop Appl. of Dig. Sig. Proc. to Audio and Acoust.,* New Palz, NY, 2003, pp. 55–58.

[4] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[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 Int. Conf. Acoust., Speech, and Sig. Proc. (ICASSP'04),* Montreal, Canada, 2004, pp. 297–300.

[6] V. R. Algazi, R. O. Duda, and D. P. Thompson, "The CIPIC HRTF database," in *Proc. IEEE Workshop Appl. of Dig. Sig. Proc. to Audio and Acoust.,* New Palz, NY, 2001, pp. 99–102.

[7] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[8] R. S. Woodworth, *Experimental Psychology.* New York: Holt, 1954.

[1] http://dept-info.labri.fr/~sm/DAFx06/