Maximum a Posteriori Binary Mask Estimation for Underdetermined Source Separation Using Smoothed Posteriors

Maximo Cobos, *Member, IEEE,* and Jose J. Lopez, *Senior Member, IEEE*

***Abstract*—Sound source separation has become a topic of intensive research in the last years. The research effort has been specially relevant for the underdetermined case, where a considerable number of sparse methods working in the time-frequency (T-F) domain have appeared. In this context, although binary masking seems to be a preferred choice for source demixing, the estimated masks differ substantially from the ideal ones. This paper proposes a Maximum a Posteriori (MAP) framework for binary mask estimation. To this end, class-conditional source probabilities according to the observed mixing parameters are modeled via ratios of dependent Cauchy distributions while source priors are iteratively calculated from the observed histograms. Moreover, spatially smoothed posteriors in the T-F domain are proposed to avoid noisy estimates, showing that the estimated masks are closer to the ideal ones in terms of objective performance measures.**

***Index Terms*—Blind Source Separation, Time-Frequency Masking, Sparse Models.**

I. Introduction

THE task of estimating and recovering independent source signals from a set of mixtures in one or several observation channels is known as *Blind Source Separation*. In the linear complete case, when as many observations as sources are available, *Independent Component Analysis* approaches are usually applied [1]. These algorithms commonly assume statistical independence and non-Gaussianity of the sources to estimate a demixing matrix that makes it possible to recover the source signals up to a permutation and scaling factor. When there are more sources than observation channels, the problem is *underdetermined* (or degenerate), and other properties such as source sparsity are exploited. Sparsity and overcomplete dictionaries have been discussed in the literature with the aim of giving a solution to the underdetermined problem, using MAP estimation [2] and l_1-norm minimization [3]. When dealing with speech and audio mixtures, it has been shown that they are sparser in the time-frequency (T-F) domain than in the time domain [4]. In fact, it has been shown that sources are almost disjoint in this domain, i.e., there exists only one source in a given T-F point. This assumption leads to the *time-frequency masking* separation approach [5]. Algorithms based on T-F masking have shown to provide significant results [6], being the ideal binary mask a commonly used benchmark for separation performance [7].

This paper proposes a MAP estimation framework for T-F masking stereo separation. In this context, two novel features are introduced to estimate the binary masks: a class-conditional distribution model for the observed anechoic mixing parameters and the use of spatially smoothed posteriors in the T-F domain. A MAP decision rule is applied to obtain the final separation masks, which are shown to provide results that are closer to the ones obtained by means of ideal binary masking.

The structure of the paper is as follows. Section II describes the signal model and the anechoic mixing parameters used as separation features. Section III proposes a statistical model for the observed mixing parameters assuming a dominant source condition. Section IV discusses binary mask estimation from a Bayesian perspective, proposing the use of smoothed posteriors for improved performance. Experiments and performance evaluation are in Section V, while the final conclusions are summarized in Section VI.

II. Signal Model and Mixing Parameters

Consider two anechoic mixture signals $x_m(t)$ given by

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - \tau_{mn}), \quad m = 1, 2, \tag{1}$$

where N is the number of sources, $s_n(t)$ are the time-domain source signals, a_{mn} are scalar coefficients and τ_{mn} are the source-to-sensor time delays. In matrix notation, the model takes the well-known form $\mathbf{x} = \mathbf{A} * \mathbf{s}$, with $\mathbf{x} = [x_1(t)\ x_2(t)]^T$, $\mathbf{s} = [s_1(t), \ldots, s_N(t)]^T$ and $\mathbf{A}_{mn} = a_{mn}\delta(t - \tau_{mn})$. In the *Short-Time Fourier Transform* (STFT) domain, the above model can be rewritten as

$$X_m(k, l) = \sum_{n=1}^{N} a_{mn} S_n(k, l) e^{-j\omega_k \tau_{mn}}, \quad m = 1, 2, \tag{2}$$

where k is the frequency bin index, l is the time-frame index, ω_k is the angular frequency corresponding to index k and $X_m(k, l)$ and $S_n(k, l)$ are the STFT versions of $x_m(t)$ and $s_n(t)$, respectively. If the sources are located in the far field, plane-wave propagation can be assumed and inter-sensor time delays τ_n are related to the *Direction-Of-Arrival* (DOA) of the sources as follows [8]

$$\tau_n = \tau_{2n} - \tau_{1n} = \frac{d}{c}\cos(\theta_n), \tag{3}$$

M. Cobos is with the Computer Science Department of the Universitat de València, 46100 Burjassot, Valencia. J.J. Lopez is with the Institute of Telecommunications and Multimedia Applications (iTEAM), Universitat Politècnica de València, 46022 Valencia, Spain. e-mail: Maximo.Cobos@uv.es (http://www.uv.es/macose2)

where d is the inter-microphone distance, c is the speed of sound and θ_n is the DOA angle of the n-th source.

A. Magnitude Ratio and DOA

Without loss of generality, the mixing process can be described by the set of amplitude mixing coefficients a_{mn} and the inter-sensor time delays τ_n resulting from the DOA of the sources. To estimate these quantities, most algorithms analyze channel differences in a sparse domain [5],[9],[8],[10],[11]. Since audio source signals do not significantly overlap in the STFT domain (a property often referred to as *W-Disjoint Orthogonality*), it can be assumed that the magnitude ratio of observation points is close (ideally equal) to the ratio of amplitude mixing coefficients:

$$R(k,l) = \arctan\left(\frac{|X_2(k,l)|}{|X_1(k,l)|}\right) \approx \arctan\left(\frac{a_{2\tilde{n}(k,l)}}{a_{1\tilde{n}(k,l)}}\right), \quad (4)$$

where $\tilde{n}(k,l)$ is the index of the dominant source at T-F point (k,l). Using the arctangent function is useful for mapping the observed values to the range $[0, \pi/2]$. On the other hand, the phase difference between mixture channels can be analyzed at each T-F element to obtain

$$D(k,l) = \frac{c}{\omega_k d}\angle\left(\frac{X_2(k,l)}{X_1(k,l)}\right) \approx \cos(\theta_{\tilde{n}(k,l)}), \quad (5)$$

where $\angle()$ denotes the phase of a complex number. Note that, according to Eq.(3), $D(k,l)$ is an estimate of the cosine of the DOA of the dominant source at point (k,l). In the following, the sources are assumed to have a unique pair of mixing parameters $(R_n,\, D_n)$ that characterizes their mixing process.

An example R-D histogram is depicted in Figure 1(a), showing the joint distribution of the mixing parameters for a mixture of 3 speech sources in a noise-free anechoic environment with $d = 2$ cm. The histogram is normalized to unit area to resemble a probability density function (pdf). The mixing parameters for each source are $(R_1, D_1) = (1.04, 0.97)$, $(R_2, D_2) = (0.40, -0.14)$ and $(R_3, D_3) = (0.32, 0.71)$. Note that the peaks in the histogram correspond to the real mixing parameters R_n and D_n, showing the presence of the different sources. The closer a point $(R(k,l),\, D(k,l))$ is to any of these peaks, the higher the chance of being dominated by the corresponding source. The rest of this paper assumes that the mixing parameters corresponding to $\mathbf{A}$ are estimated from these peaks. In fact, recent studies have shown that, while mixing matrix estimation for non-reverberant underdetermined mixtures is a task that can be successfully accomplished, the unmixing procedure is still the main challenge [7].

III. Modeling of Mixing Parameters

Sparse signals are usually modeled by distributions having sharp peaks at zero and flat tails. The Cauchy (or Lorentz) distribution, $\mathcal{C}(\mathsf{x}_0, \gamma)$, describes properly magnitude sparsity due to its peaky and heavy-tailed nature, accounting for rarely appearing high values [12]. Its probability density function is given by

$$f(\mathsf{x}) = \frac{1}{\pi}\left[\frac{\gamma}{(\mathsf{x}-\mathsf{x}_0)^2+\gamma^2}\right], \quad (6)$$

where x_0 specifies the peak location of the distribution and γ the half-width at half-maximum. To statistically model the joint distribution of R and D, first, the STFT of the sources $S_n(k,l)$ are assumed to be independent complex random processes as follows:

$$|S_n(k,l)| \sim \beta_n|\mathcal{C}(0,1)|, \quad (7)$$

$$\angle S_n(k,l) \sim \mathcal{U}(-\pi,\pi), \quad (8)$$

where $\mathcal{C}(0,1)$ denotes samples drawn from a normalized Cauchy distribution centered at zero with $\gamma = 1$, while $\mathcal{U}(-\pi,\pi)$ refers to a uniform distribution in the range $[-\pi,\pi]$. The parameter β_n represents the relative contribution of the n-th source . As a result, the source model, expressed by $\mathcal{S}_n$, is written

$$\mathcal{S}_n \sim \beta_n|\mathcal{C}(0,1)|e^{j\mathcal{U}(-\pi,\pi)}, \quad n = 1,\ldots,N. \quad (9)$$

The goal now is to obtain the distribution of the mixing parameters for points where a given source in a mixture is dominant. The mixture magnitude ratio, according to Eq.(4) and assuming unit-norm mixing matrix columns, is given by

$$\mathcal{R}^{\mathbf{A}\boldsymbol{\beta}} = \arctan\left(\frac{\left|\sum_{n=1}^{N}\sin(R_n)\mathcal{S}_n e^{-j((d/c)\breve{\omega}D_n)}\right|}{\left|\sum_{n=1}^{N}\cos(R_n)\mathcal{S}_n\right|}\right), \quad (10)$$

where $\breve{\omega} \in \omega_k$ is a random frequency value. Similarly, the observed DOAs are given by

$$\mathcal{D}^{\mathbf{A}\boldsymbol{\beta}} = \frac{c}{2\pi\breve{\omega}d}\angle\left(\frac{\sum_{n=1}^{N}\sin(R_n)\mathcal{S}_n e^{-j((d/c)2\pi\breve{\omega}D_n)}}{\sum_{n=1}^{N}\cos(R_n)\mathcal{S}_n}\right). \quad (11)$$

The use of $\mathbf{A}$ and $\boldsymbol{\beta}$ in the notation of $\mathcal{R}^{\mathbf{A}\boldsymbol{\beta}}$ and $\mathcal{D}^{\mathbf{A}\boldsymbol{\beta}}$ denotes its dependence on the estimated mixing parameters and the selected parameter vector $\boldsymbol{\beta} = [\beta_1,\, \beta_2\, \ldots, \beta_N]^T$. Note that both $\mathcal{R}^{\mathbf{A}\boldsymbol{\beta}}$ and $\mathcal{D}^{\mathbf{A}\boldsymbol{\beta}}$ are simulated values obtained from the $\mathcal{S}_n$ random data.

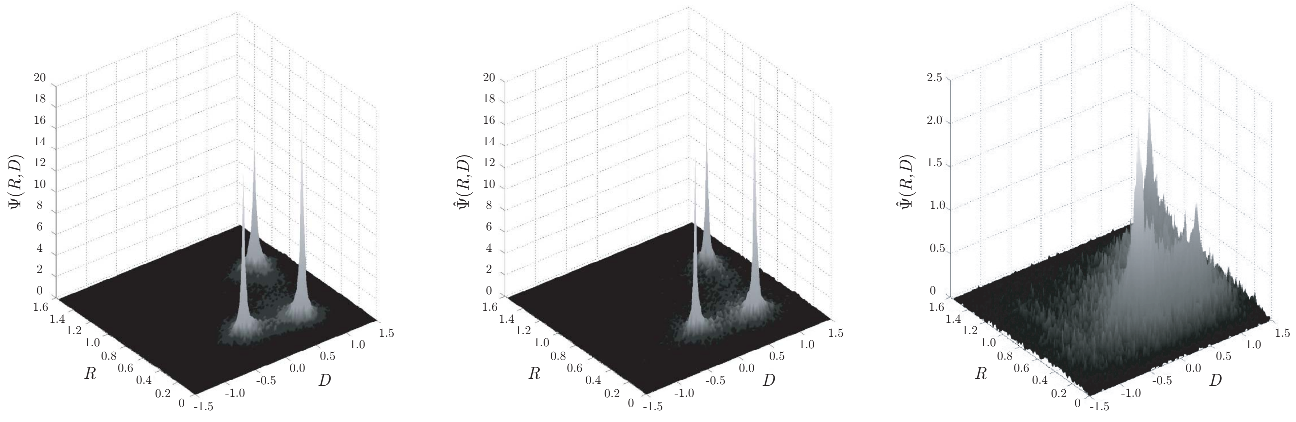
The magnitude ratio and DOA distributions for a dominant source can be extracted from the above by taking the set of points where its magnitude is dominant over the rest:

$$\mathcal{R}_n^{\mathbf{A}\boldsymbol{\beta}} = \left\{\mathcal{R}^{\mathbf{A}\boldsymbol{\beta}} \in |\mathcal{S}_n| > \sum_{n' \neq n}|\mathcal{S}_{n'}|\right\}, \quad n = 1,\ldots,N, \quad (12)$$

$$\mathcal{D}_n^{\mathbf{A}\boldsymbol{\beta}} = \left\{\mathcal{D}^{\mathbf{A}\boldsymbol{\beta}} \in |\mathcal{S}_n| > \sum_{n' \neq n}|\mathcal{S}_{n'}|\right\}, \quad n = 1,\ldots,N. \quad (13)$$

Since obtaining closed-form expressions for dominant source distributions is very difficult, the use of Monte-Carlo processing for their computation offers a practical solution. The only free parameters of the model are the β_n, which are easily determined numerically by iteratively fitting the distribution of the simulated data to the one of the real mixture histogram $\Psi(R,D)$. To this end, the difference between peak amplitudes in both histograms is minimized in the least squares sense. Thus, we define an observed peak amplitude vector $\mathbf{b}(\boldsymbol{\beta}) = [b_1,\, b_2, \ldots, b_N]^T$ and a target peak amplitude vector $\mathbf{p} = [p_1,\, p_2, \ldots, p_N]^T$, with

$$b_n = \hat{\Psi}\left(R_n, D_n\right), \quad n = 1,\ldots,N, \quad (14)$$



(a) (b) (c)

Fig. 1. R-D histograms for the example mixture. (a) Real histogram. (b) Histogram obtained by numerical processing using a Cauchy source model. (c) Histogram using a Laplacian source model.

$$p_n = \Psi\left(R_n, D_n\right), \quad n = 1, \ldots, N, \tag{15}$$

where $\hat{\Psi}(R,D)$ is the normalized R-D histogram (unit area) computed from the $\mathcal{R}^{\mathbf{A}\boldsymbol{\beta}}$ and $\mathcal{D}^{\mathbf{A}\boldsymbol{\beta}}$ synthetic data. The parameter vector $\boldsymbol{\beta}$ is iteratively updated until convergence as follows (see Appendix):

$$\boldsymbol{\beta}^{+} = \boldsymbol{\beta} - \eta\left(\bar{\mathbf{b}}(\boldsymbol{\beta}) - \bar{\mathbf{p}}\right), \tag{16}$$

where $\bar{\mathbf{b}}(\boldsymbol{\beta}) = \frac{\mathbf{b}(\boldsymbol{\beta})}{||\mathbf{b}(\boldsymbol{\beta})||}$ and $\bar{\mathbf{p}} = \frac{\mathbf{p}}{||\mathbf{p}||}$ are the generated and target normalized peak amplitudes, respectively. Figure 1(b) shows the histogram calculated by means of the proposed model as a sum of the individual dominant source distributions. Note that it is very similar to the one of Figure 1(a), i.e. the extracted from the real mixture.

A. Suitability of the Model

Source sparsity in the T-F domain has been discussed in many works. Statistical models for source distributions in sparse domains are usually based on super-Gaussian distributions, such as the Laplacian distribution [13]. In fact, it has been reported that the Laplacian distribution is able to model speech both in the time domain and in the STFT domain [14]. Moreover, due to the peaky nature of the magnitude ratio distribution, Laplacian mixture models have also been employed in underdetermined blind source separation problems [15]. However, while the Laplacian distribution might be suitable to model STFT coefficients under some circumstances [16], the proposed Cauchy-based model for STFT coefficients has been shown to provide better accuracy in our R-D modeling task. Both source models have been compared by means of a χ^2 test to evaluate their capability to generate synthetic data fitting real R-D distributions. The value of the test statistic is given by

$$\chi^2 = \frac{\sum_R \sum_D (n_{R,D} - N_p \Delta_R \Delta_D \Psi(R,D))^2}{N_p \Delta_R \Delta_D \Psi(R,D)}, \tag{17}$$

where N_p is the total number of T-F points in the mixture, $n_{R,D}$ is the number of generated sample points with D and R falling into a given interval of the histogram and Δ_R and Δ_D are the lengths of a histogram bin in the R and D axis, respectively. The smaller the χ^2 value, the better the fit. The test was performed over the anechoic test signals used in Section V, resulting in a mean value of $\chi^2 = 4.8e^4$ for the proposed model and $\chi^2 = 1.81e^5$ for the Laplacian model. The R-D distribution obtained for the example mixture using a Laplacian model is shown in Figure 1(c). Note that the proposed model (b) is significantly closer to the real distribution.

IV. Separation Mask Estimation

The statistical model described in the previous section allows to compute a class-conditional probability measure for the mixing parameters R, D given a dominant source. The likelihoods are therefore given by

$$p(R,D|s_n) = \hat{\Psi}_n(R,D)., \quad n = 1, \ldots, N, \tag{18}$$

where $\hat{\Psi}_n(R,D)$ are the normalized histograms computed from the model data $\mathcal{R}_n^{\mathbf{A}}$, $\mathcal{D}_n^{\mathbf{A}}$. The factors β_n are considered priors for each source obtained from the whole observation time:

$$P(s_n) = \beta_n, \quad n = 1, \ldots, N, \tag{19}$$

properly scaled so that $\sum_{n=1}^{N} \beta_n = 1$. Then, the posterior probability according to Bayes' theorem is given by

$$P(s_n|R,D) = \frac{p(R,D|s_n)P(s_n)}{p(R,D)}, \quad n = 1, \ldots, N, \tag{20}$$

with $p(R,D) = \sum_{n=1}^{N} p(R,D|s_n)P(s_n)$. A first estimation of the separation masks could be obtained by applying the following MAP decision rule:

$$M_n(k,l) = \begin{cases} 1 & \text{if} \quad n = \arg\max_{n'} p(R,D|s_{n'})P(s_{n'}) \\ 0 & \text{elsewhere} \end{cases} \quad \forall (k,l). \tag{21}$$

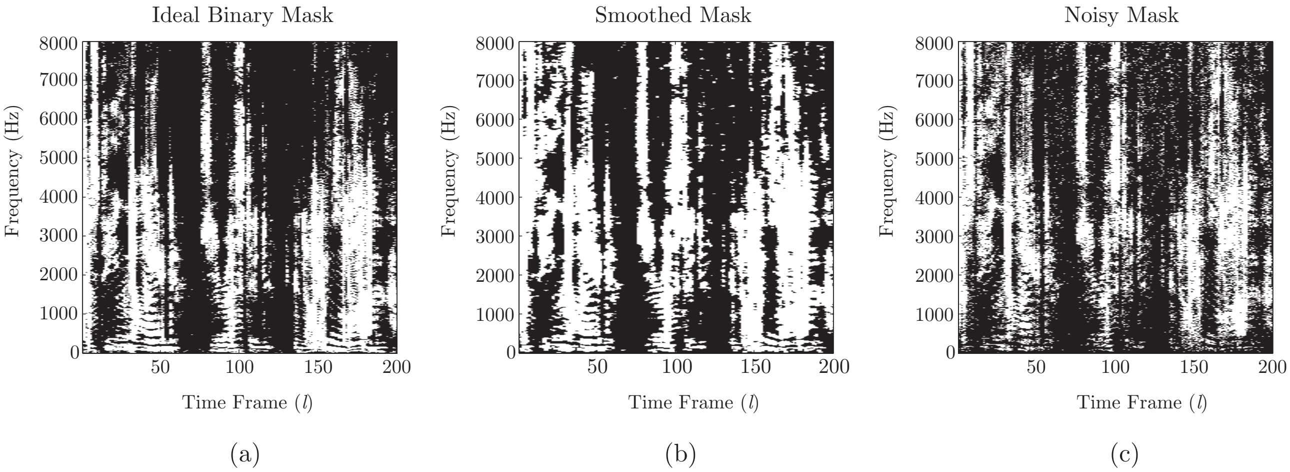


Fig. 2. Binary masks for one of the three sources of the example mixture. Reliable elements are indicated by white T-F units and unreliable T-F units are marked in black. (a) Ideal binary mask. (b) Improved mask obtained with smoothed posteriors. (c) Noisy mask obtained without smoothing.

A. Smoothed Time-Frequency Posteriors

It is well-known that ideal binary masks show clusters of points corresponding to those areas where the energy of a given source is higher than the total interfering energy. Due to the nature of speech and audio sources, these clusters usually appear distributed around speech partials and other high-energy components. Therefore, if a given T-F point of a separation mask is active, it is likely that surrounding points are also active. Similarly, zero points corresponding to silences or low-energy areas may be surrounded also by other inactive points. This property is exploited in this section by using smoothed T-F posteriors as follows. Let us introduce in our decision the evidence of observing the probability that T-F points pertaining to a surrounding neighborhood $\Omega_{k,l}$ do also belong to the same source:

$$P(s_n|R,D,\Omega_{k,l}) = \frac{p(R,D,\Omega_{k,l}|s_n)P(s_n)}{p(R,D,\Omega_{k,l})}. \qquad (22)$$

Obviously, computing the new likelihood $p(R,\Omega_{k,l}|s_n)$ for all possible neighborhoods is not very practical, even if conditional independence is assumed. However, it seems reasonable to think that the belief given by the posterior will change accordingly to the support of surrounding points due to their existing correlation. Thus, a convolution operation is proposed to model this influence:

$$P(s_n|R,D,\Omega_{k,l}) = P(s_n|R,D) * W(k,l), \qquad (23)$$

where $W(k,l)$ is a two-dimensional smoothing impulse response, implemented by a properly normalized smoothing matrix (kernel). Smoothed posteriors have been widely employed in image processing for pixel classification [17],[18]. We propose the use of a Gaussian filter, which gives more importance to the central point but smoothly incorporates information from the surrounding points. Then, the suggested impulse response as a function of time t (in ms) and frequency f (in Hz) is expressed as

$$W(f,t) = \frac{1}{2\pi\sigma_f\sigma_t} e^{-\left(\frac{f^2}{2\sigma_f^2}+\frac{t^2}{2\sigma_t^2}\right)}, \qquad (24)$$

where σ_f^2 and σ_t^2 are the variances of the Gaussian filter that control the area of influence $\Omega_{k,l}$ in the frequency and time dimensions, respectively. Note that in in the above definition, the variables k and l have been replaced for t and f in order to make the selected filter independent of the STFT analysis parameters. Additionally, a T-F invariant smoothing kernel has been proposed for the sake of simplicity. However, it is worth to note that ideal binary masks have more horizontal structure in low frequencies and more vertical structure in high frequencies. Although frequency dependent kernels might benefit the mask estimation task, the study of different kernel alternatives is out of the scope of this paper and will be addressed in a future work.

The final separation masks are obtained by applying the MAP decision rule over the smoothed posteriors:

$$M_n(k,l) = \begin{cases} 1 & \text{if} \quad n = \arg\max_{n'} P(s_{n'}|R,D,\Omega_{k,l}) \\ 0 & \text{elsewhere} \end{cases} \quad \forall(k,l). \qquad (25)$$

The estimated sources are recovered by applying the estimated mask to the mixture channels and transforming the signals back to the time-domain with the inverse STFT operator. Figure 2(a)-(c) compares visually an ideal binary mask corresponding to one of the sources of the example mixture with the ones obtained with Eq.(25) and Eq.(21). In the next section, we evaluate the separation performance of the proposed method in terms of objective performance measures.

V. Experiments and Evaluation

In this section, a performance evaluation is presented in terms of the well-known objective performance measures *Signal to Distortion Ratio* (SDR), *source Image Spatial distortion Ratio* (ISR), *Source to Interference Ratio* (SIR) and *Source to Artifacts Ratio* (SAR) [19].

The proposed method was evaluated and compared to other separation approaches using underdetermined mixtures of $N = 3$ and $N = 4$ sources. The source signals were male and female speech fragments ($f_s = 16$ kHz) provided with the

TABLE I
AVERAGE PERFORMANCE IN ANECHOIC SCENARIO

	$N = 3$ sources							$N = 4$ sources					
	Prop	*Prop w/s*	*DUET*	*S-DUET*	*Duong*	*IBM*		*Prop*	*Prop w/s*	*DUET*	*S-DUET*	*Duong*	*IBM*
SDR	7.23	5.12	4.55	4.36	6.11	10.64	SDR	4.79	3.62	2.96	2.77	3.25	9.22
SIR	17.34	17.32	11.47	14.44	10.32	21.58	SIR	12.87	12.79	8.26	11.72	4.82	19.55
SAR	7.46	5.29	5.10	4.65	9.43	11.07	SAR	4.76	3.31	3.18	2.36	6.73	9.61
ISR	14.22	10.23	8.44	9.78	11.35	20.08	ISR	10.45	7.44	6.35	6.21	6.81	17.63

TABLE II
AVERAGE PERFORMANCE IN REVERBERANT SCENARIO ($T_{60} = 250\ ms$)

	$N = 3$ sources							$N = 4$ sources					
	Prop	*Prop w/s*	*DUET*	*S-DUET*	*Duong*	*IBM*		*Prop*	*Prop w/s*	*DUET*	*S-DUET*	*Duong*	*IBM*
SDR	3.61	3.56	3.26	3.46	4.40	10.76	SDR	2.12	2.04	1.94	1.96	2.44	9.29
SIR	6.87	6.69	5.11	6.46	5.00	20.33	SIR	4.00	3.88	3.52	4.32	1.61	19.45
SAR	5.50	5.15	5.10	4.72	7.78	11.01	SAR	3.19	3.36	3.62	2.47	6.02	9.73
ISR	9.36	8.74	8.15	7.97	10.05	19.84	ISR	7.11	6.43	6.65	5.22	5.95	17.47

'*Dev2*' dataset of the *Signal Separation Evaluation Campaign* (SiSEC) [7]. To evaluate the influence of room reflections, two simulated scenarios were considered: an anechoic environment and a box-shaped room ($5 \times 4 \times 3$ m) with reverberation time $T_{60} = 250$ ms. The inter-microphone distance was $d = 0.2$ cm to avoid possible spatial aliasing effects. The sources were randomly positioned at different directions to generate a total of 50 test mixtures in the anechoic scenario and 50 test mixtures in the reverberant one. STFTs were computed using Hamming windows of 1024 samples length and 75% overlap. The proposed method was applied using a Gaussian smoothing filter with variances $\sigma_t^2 = 25$ ms and $\sigma_f^2 = 23$ Hz, implemented by a 3×3 kernel matrix.

The following systems were compared: the proposed approach (*Prop*), the proposed approach without using smoothed posteriors (*Prop w/s*), the DUET algorithm [5] (*DUET*), the smoothed DUET algorithm using a plus sign-shaped median filter [11] (*S-DUET*), the algorithm for reverberant mixtures by Duong et al. [20] (*Duong*) and ideal binary masking (*IBM*). The ideal binary mask was computed by comparing the target source signal and the interfering source images as in [7].

Results for the anechoic and reverberant environments are shown in Table I and Table II, respectively. It can be observed that, in the anechoic environment, the proposed method outperforms all the other systems, providing results that are closer to the ones obtained by ideal binary masking. Moreover, the usefulness of the model is here demonstrated in terms of separation performance, since the results obtained by our method without smoothing are still better than the ones of DUET and S-DUET. In the case of reverberant mixtures, the proposed method still provides better results than DUET and S-DUET but only outperforms Duong's algorithm in terms of SIR and ISR. This fact highlights the importance of having a suitable model for a specific application scenario since, as opposed to Duong's algorithm, the proposed method does not take into account room reverberation effects. Nevertheless, note that, as the number of sources increases, our proposed method and Duong's tend to be comparable as shown by the results with $N = 4$. Despite the proposed method has been shown to be the most effective in anechoic scenarios, further work would be needed to make it more robust to room reflections.

VI. CONCLUSION

This paper presented a T-F masking separation method developed from a MAP perspective. Two main novel features were introduced with respect to other T-F masking approaches. First, a likelihood model for the observed mixing parameters under a source dominance assumption was described. To this end, ratios of complex dependent Cauchy distributions were computed and statistically characterized by means of Monte-Carlo processing. Second, smoothed posteriors in the MAP decision were proposed to model the influence of neighboring T-F points, reducing the amount of noisy points in the estimated masks. The proposed method was shown to outperform other separation approaches in anechoic and reverberant environments, providing average results closer to the ones provided by the ideal binary masking benchmark. However, further work is needed to adapt the model to a reverberant case to make it more robust against room reflections.

APPENDIX

The optimum parameter vector $\boldsymbol{\beta}$ is found by iteratively minimizing the difference between the distributions of the real measured data and the synthetically generated data. To this end, we search for the best fit in the least squares sense by considering only the relative amplitude of the peaks (normalized to unit power) in both histograms. The solution is found by minimizing the function

$$F(\boldsymbol{\beta}) = \left\|\bar{\mathbf{b}}(\boldsymbol{\beta}) - \bar{\mathbf{p}}\right\|^2, \qquad (26)$$

where $\bar{\mathbf{b}}(\boldsymbol{\beta}) = \frac{\mathbf{b}(\boldsymbol{\beta})}{||\mathbf{b}(\boldsymbol{\beta})||}$ and $\bar{\mathbf{p}} = \frac{\mathbf{p}}{||\mathbf{p}||}$ are the generated and observed normalized peak amplitudes, respectively. Next, we assume a simplified linear model $\bar{\mathbf{b}}(\boldsymbol{\beta}) = \mathbf{D}\boldsymbol{\beta}$, where the amplitude of a peak $\bar{b}_n$ only depends on its corresponding β_n parameter:

$$\begin{bmatrix} \bar{b}_1 \\ \bar{b}_2 \\ \vdots \\ \bar{b}_N \end{bmatrix} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & d_{NN} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}. \qquad (27)$$

Thus, considering the above diagonal matrix, the gradient of $F(\boldsymbol{\beta})$ can be expressed as

$$\nabla F(\boldsymbol{\beta}) = 2\mathbf{D}\left(\bar{\mathbf{b}}(\boldsymbol{\beta}) - \bar{\mathbf{p}}\right). \quad (28)$$

Moreover, we can further assume that all the peaks have the same linear dependence with its associated parameter ($\mathbf{D} = \alpha\mathbf{I}$), leading to the next simplified gradient expression

$$\nabla F(\boldsymbol{\beta}) = 2\alpha\left(\bar{\mathbf{b}}(\boldsymbol{\beta}) - \bar{\mathbf{p}}\right). \quad (29)$$

Finally, the optimized $\boldsymbol{\beta}$ parameters can be iteratively found by following a gradient descent approach:

$$\boldsymbol{\beta}^{+} = \boldsymbol{\beta} - \gamma\nabla F(\boldsymbol{\beta}) = \boldsymbol{\beta} - \eta\left(\bar{\mathbf{b}}(\boldsymbol{\beta}) - \bar{\mathbf{p}}\right), \quad (30)$$

where $\gamma > 0$ is a small number controlling the step size. The constant factor $\eta = 2\gamma\alpha$ can be experimentally adjusted. In this paper, we assumed $\eta = 1$.

Acknowledgment

The Spanish Ministry of Science and Innovation supported this work under the project TEC2009-14414-C03-01.

References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Oxford, UK: Academic Press, 2010.

[2] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.

[3] D. L. Donoho and M. Elad, "Maximal sparsity representation via l_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 2197–2202, 2003.

[4] J. J. Burred and T. Sikora, "On the use of auditory representations for sparsity-based sound source separation," in *Proceedings of the 5th International Conference on Information, Communications and Signal Processing (ICICS 2005)*, Bangkok, Thailand, December 2005.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[6] D. Wang, "Time frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[7] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," *Lecture Notes in Computer Science. Independent Component Analysis and Signal Separation*, vol. 5441/2009, pp. 734–741, 2009.

[8] M. Cobos and J. J. Lopez, "Two-microphone separation of multiple speakers based on interclass variance maximization," *Journal of the Acoustical Society of America*, vol. 127, pp. 1661–1673, 2010.

[9] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[10] B. Gunel, H. Hacihabiboglu, and A. M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 4, pp. 748–756, 2008.

[11] M. Kühne, R. Togneri, and S. Nordholm, *Speech Recognition, Technologies and Applications*. I-Tech, 2008, ch. Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition, pp. 61–80.

[12] S. Kotz, N. Balakrishnan, C. B. Read, and B. Vidakovic, Eds., *Enclyclopedia of Statistical Sciences (2nd Edition)*. John Wiley & Sons, 2005, vol. 2.

[13] H. Rabbani and S. Gazor, "Local probability distribution of natural signals in sparse domains," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, 2011, pp. 1289–1292.

[14] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

[15] N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using laplacian mixture models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1818–1832, 2007.

[16] R. Martin and C. Breithaupt, "Speech enhancement in the DFT domain using laplacian speech priors," in *International Workshop on Acoustic Echo and Noise Control (IWAENC 2003)*, Kyoto, Japan, 2003, pp. 87–90.

[17] P. Teo, G. Sapiro, and B. Wandell, "Anisotropic smoothing of posterior probabilities," in *IEEE International Conference on Image Processing*, Santa Barbara, CA, USA, 1997, pp. 675–678.

[18] S. Haker, G. Sapiro, and A. Tannenbaum, "Knowledge-based segmentation of SAR data with learned priors," *IEEE Transactions on Image Processing*, vol. 9, no. 2, pp. 299–301, 2000.

[19] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

Maximo Cobos was born in Alicante, Spain, in 1982. He received a telecommunications engineer degree in 2006, a M.S. degree in telecommunication technologies in 2007 and the Ph.D degree in telecommunications in 2009, all of them from the Universitat Politècnica de València, Spain. His Ph.D dissertation was awarded with the Ericsson Best Thesis Award on Multimedia Environments from the Spanish National Telecommunications Engineering Association (COIT). He completed with honors his Ph.D studies under the University Faculty Training program (FPU). In 2010 he was awarded with a 'Campus of Excellence' post-doctoral fellowship to work at the Institute of Telecommunications and Multimedia Applications (iTEAM). In 2009 and 2011, he was a visiting researcher at Deutsche Telekom Laboratories in Berlin, where he worked in the field of audio signal processing for telecommunications. Since 2011 he is an Assistant Professor at the Computer Science Department of the Universitat de València. His work is focused on the area of digital signal processing for audio and multimedia applications, where he has published more than 50 technical papers in international journals and conferences. Dr. Cobos is a member of the IEEE and the Audio Engineering Society (AES).

Jose J. Lopez was born in Valencia, Spain, in 1969. He received a telecommunications engineer degree in 1992 and a Ph.D. degree in 1999, both from the Universitat Politècnica de València, Spain. Since 1993 he has been involved in education and research at the Communications Department of the Universitat Politècnica de València, where he is currently a Full Professor. His research activity is centered on digital audio processing in the areas of spatial audio, wave field synthesis, physical modeling of acoustic spaces, efficient filtering structures for loudspeaker correction, sound source separation and development of multimedia software in real time. Dr. Lopez has published more than 160 papers in international technical journals and at renowned conferences in the fields of audio and acoustics and has led more than 25 research projects. He was workshop cochair at the 118th Convention of the Audio Engineering Society in Barcelona and has been serving on the committee of the AES Spanish Section for 9 years, at present as secretary of the Section. He is a senior member of the IEEE, a full ASA member and an AES member.