2012 2nd International Conference on Computer Design and Engineering (ICCDE 2012)
IPCSIT vol. 49 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V49.16

Magnitude Ratio Modelling of Instantaneous Stereo Audio Mixtures in the Time-Frequency Domain

Maximo Cobos[1+], Jaume Segura[1] and Jose J. Lopez[2]

[1] Departamento de Informática, Universitat de València
[2] iTEAM Institute, Universitat Politècnica de València

Abstract. Underdetermined sound source separation has become a topic of intensive research in the last years. The separation of audio sources from an underdetermined stereo mixture can be relevant in many applications, such as music transcription, object-based audio coding, audio remixing or music information retrieval tasks. There is a considerable number of methods aimed at solving the underdetermined instantaneous mixing problem, many of them based on the sparsity provided by time-frequency (T-F) representations. Moreover, the magnitude ratio of the mixture channels in the time-frequency domain is often employed as an important feature in the separation process. In this paper, we study the magnitude ratio distribution in the time-frequency domain for underdetermined stereo mixtures and propose a likelihood model for the magnitude ratio under a source dominance assumption.

Keywords: Source Separation, Magnitude Ratio Modelling, Audio Signal Processing, Time-Frequency Processing.

1. Introduction

Blind Source Separation is the task of estimating and recovering independent source signals from a set of mixtures in one or several observation channels. In the linear complete case, when as many observations as sources are available, Independent Component Analysis (ICA) approaches are usually applied [1]. Algorithms based on ICA usually 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 maximum a posteriori estimation [2] and l1-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. Algorithms based on T-F masking have shown to provide significant results [5], being the ideal binary mask a commonly used benchmark for separation performance [6].

The magnitude ratio of mixture channels in the T-F domain has been shown to be a meaningful feature in the design of sound source separation algorithms. This ratio is directly related to the spatial arrangement of the sources in the stereo mixture and, thus, a statistical model for this parameter can be very useful to develop high-performance separation methods. This paper proposes the use of monte-carlo processing to compute a class-conditional probability model for the magnitude-ratio of underdetermined stereo audio mixtures. Moreover, a Maximum-A-Posteriori (MAP) decision rule will be derived to show how the proposed model can be effectively used in binary-masking source separation. The structure of the paper is as

+ Corresponding author. *E-mail address:* maximo.cobos@uv.es.

follows. Section 2 describes the signal model and defines the magnitude ratio as a separation feature. Section 3 proposes a statistical model for the observed magnitude ratio assuming a dominant source condition. Section 4 discusses binary mask estimation from a Bayesian perspective, proposing the use of smoothed posteriors for improved performance. Experiments and performance evaluation are in Section 5, while the final conclusions are summarized in Section 6.

2. Signal Model and Magnitude Ratio

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

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

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

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

where k is the frequency bin index, l is the time-frame index and $X_m(k,l)$ and $S_n(k,l)$ are the STFT versions of $x_m(t)$ and $s_n(t)$, respectively.

2.1. Magnitude Ratio and Estimation of Mixing Coefficients

To estimate the mixing coefficients a_{mn}, most algorithms analyze channel differences in a sparse domain [5],[7],[8],[9]. 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 mixing coefficients:

$$R(k,l) = \arctan\left(\frac{|X_2(k,l)|}{|X_1(k,l)|}\right) \approx \arctan\left(\frac{a_{2\tilde{n}(k,l)}}{a_{1\tilde{n}(k,l)}}\right), \tag{3}$$

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, π/2]. An example histogram of R, denoted as $\Psi(R)$, is shown in is shown in Fig.1(b). The peaks in the histogram correspond to the real mixing ratios, showing the presence of the different sources (4 speech sources in the example). The closer a point $R(k, l)$ 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 coefficients are easily estimated (up to a scaling and permutation) from these peaks, so that **A** is known. Fig.1(b) also shows the contribution of each source to the total histogram by depicting the distribution of the magnitude ratio at those T-F points where each source is dominant, i.e., points corresponding to the ideal binary masks. In the following section, we model these contributions by taking ratios of dependent Cauchy distributions. It will be shown how a class-conditional probability measure can be derived for each source given the observed magnitude ratio.

3. Magnitude Ratio Modeling

Source sparsity in the T-F domain has been discussed in many works. Statistical models for source distributions are usually based on super-Gaussian distributions, such as the Laplacian distribution [10]. While these models are sufficient for many purposes, a more accurate model can be obtained by assuming sources with STFT coefficients having Cauchy-distributed magnitude and uniformly distributed phase. The Cauchy (or Lorentz) distribution, $\mathcal{C}(\mathrm{x}_0, \gamma)$, describes properly magnitude sparsity due to its peaky nature and its very flat tails, accounting for rarely appearing high values [11]. Its probability density function is given by

$$f(\mathrm{x}) = \frac{1}{\pi}\left[\frac{\gamma}{(\mathrm{x}-\mathrm{x}_0)^2+\gamma^2}\right], \tag{4}$$

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

$$\begin{aligned} |S_n(k,l)| &\sim \beta_n \mathcal{C}(0,1), \\ \angle S_n(k,l) &\sim \mathcal{U}(-\pi,\pi), \end{aligned} \tag{5}$$

where $\mathcal{C}(0,1)$ C denotes samples drawn from a normalized Cauchy distribution centered at zero with γ = 1, while $\mathcal{U}(-\pi,\pi)$ refers to a uniform distribution in the range [-π, π]. 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

$$\begin{aligned} |S_n(k,l)| &\sim \beta_n \mathcal{C}(0,1), \\ \angle S_n(k,l) &\sim \mathcal{U}(-\pi,\pi), \end{aligned} \tag{6}$$

The goal now is to obtain the distribution of the magnitude ratio for points where a given source in a mixture is dominant. This is not an easy task, since it implies ratios of dependent linear combinations of the above defined distributions. The mixture magnitude ratio, according to Eq.(3), is given by

$$\mathcal{R}^{\mathbf{A}} = \arctan\left(\frac{|\sum_{n=1}^{N} a_{2n}\widetilde{\mathcal{S}}_n|}{|\sum_{n=1}^{N} a_{1n}\widetilde{\mathcal{S}}_n|}\right). \tag{7}$$

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

$$\mathcal{R}_n^{\mathbf{A}} = \left\{\mathcal{R}^{\mathbf{A}} \in |\mathcal{S}_n| > \sum_{n'\neq n} |\mathcal{S}_{n'}|\right\}, \quad n = 1,\ldots,N. \tag{8}$$

Note that obtaining closed-form expressions for dominant source ratio distributions is very difficult and out of the scope of this paper, being a more practical solution to use Monte-Carlo processing for their computation. The only free parameters of the model are the β_n, which are easily determined numerically. To this end, they are firstly initialized to the value that the observed histogram Ψ takes at the real magnitude ratios, i.e. at its local maxima $\beta_n = \Psi(a_{2n}/a_{1n})$. Then, they are iteratively updated until convergence as follows:

$$\begin{aligned} \beta_n^{+} &= \beta_n + e_n, \quad n = 1,\ldots,N, \\ e_n &= \Psi\left(\arctan\left(\frac{a_{2n}}{a_{1n}}\right)\right) - \hat{\Psi}\left(\arctan\left(\frac{a_{2n}}{a_{1n}}\right)\right), \end{aligned} \tag{9}$$

where $\hat{\Psi}(R)$ is the normalized histogram (unit area) computed from the $\mathcal{R}^{\mathbf{A}}$ synthetic data. Fig.1(a) shows the histograms obtained assuming the mixing coefficients of the example mixture. The real histograms corresponding to the example are shown in Fig.1(b). Note that the histograms in (b) can only be obtained if the sources are previously known, here residing the usefulness of the proposed model.

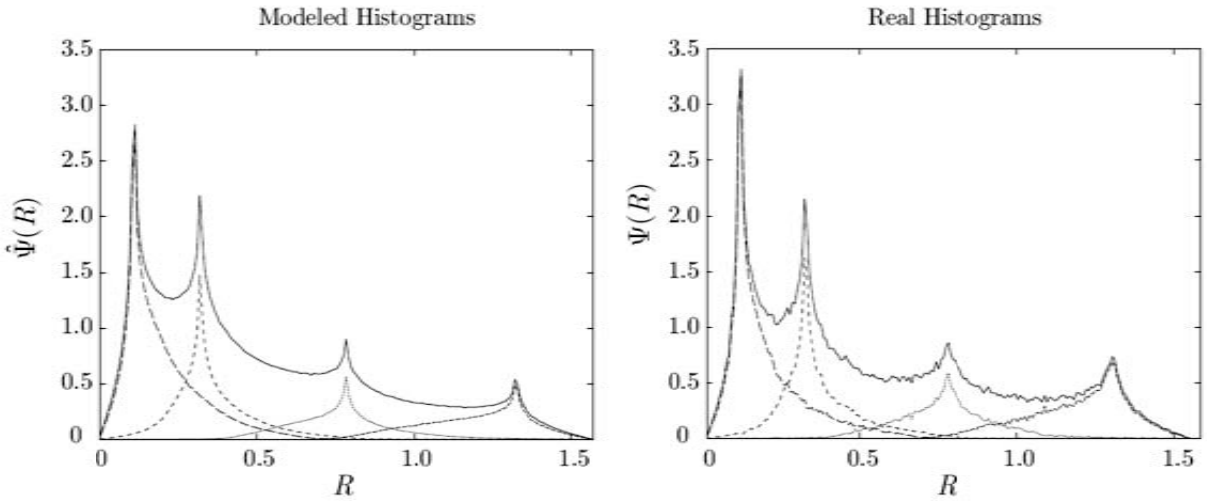

Fig. 1: Histograms of the magnitude ratio for the example mixture. (a) Model obtained by numerical processing. (b) Real histograms.

4. Source Separation

The statistical model described in the previous section allows to compute a class-conditional probability measure for the magnitude ratio *R* given a dominant source. The likelihoods are therefore given by

$$p(R \mid s_n) = \hat{\Psi}_n(R)., \quad n = 1, \ldots, N, \tag{10}$$

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

$$P(s_n) = \beta_n, \quad n = 1, \ldots, N, \tag{11}$$

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

$$P(s_n \mid R) = \frac{p(R \mid s_n) P(s_n)}{p(R)}, \quad n = 1, \ldots, N, \tag{12}$$

With $p(R) = \sum_{n=1}^{N} p(R \mid s_n) P(s_n)$. The separation masks are therefore obtained as

$$M_n(k,l) = \begin{cases} 1 & \text{if} \quad n = \operatorname{argmax}_{n'} p(R \mid s_{n'}) P(s_{n'}) \\ 0 & \text{elsewhere} \end{cases} \quad \forall k, l. \tag{13}$$

5. 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) [12]. Random mixtures were artificially generated by using the speech (male and female) and music source signals provided with the development database of the Signal Separation Evaluation Campaign (SiSEC), see details in [6]. A total of 350 speech mixtures (with 4 sources) and 150 music mixtures (with 3 sources) were generated and processed (fs = 16 kHz). STFTs were computed using Hann windows of 1024 samples length and 50% overlap. Results are shown in Table 1, which compares the average separation performance for DUET [5], the proposed method (Prop.) and ideal binary masking (IBM). It can be observed that the results obtained by our estimated masks outperform those estimated with DUET in all the objective performance measures, being closer to the IBM benchmark. This is true both for speech mixtures and music mixtures, showing that the proposed framework provides a valuable solution for underdetermined stereo source separation.

	Speech ($N = 4$)			Music ($N = 3$)		
	DUET	Prop.	IBM	DUET	Prop.	IBM
SDR	5,1	6,6	9,0	4,3	8,6	11,4
ISR	12,0	14,9	17,9	12,5	17,1	18,9
SIR	14,1	16,9	20,0	11,4	17,8	19,6
SAR	5,5	6,4	9,3	6,2	9,9	13,1

Table 1: Average Separation Performance.

6. Conclusion

This paper presented a time-frequency masking separation method developed from a Bayesian perspective. Two main novel features were introduced with respect to other T-F masking approaches. First, a likelihood model for the magnitude ratio under a source dominance assumption was described. To this end, ratios of complex dependent Cauchy distributions are computed and statistically characterized by means of Monte-Carlo processing. The proposed method was shown to outperform other binary masking approaches, providing average results closer to the ones provided by the ideal binary masking benchmark. Future work will be focused on introducing new features in the decision, such as energy-dependent risks or pitch-based model priors.

7. Acknowledgements

This work is supported by the Spanish Ministry of Science and Technology (MCYT) under Project ref. TEC2009-14414-C03-01 and FEDER funds.

8. 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, 2000, **12**: 337–365,.

[3] D. L. Donoho and M. Elad, "Maximal sparsity representation via l1 minimization," Proceedings of the National Academy of Sciences, 2003, **100**: 2197–2202.

[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, 2004, **52**(7): 1830–1847.

[6]] 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, 2009, **5441/2009**: 734–741.

[7] P. Bofill and M. Zibulevski, "Underdetermined blind source separation using sparse representations," Signal Processing, 2001, **81:** 2353–2362.

[8] M. Cobos and J. J. Lopez, "Stereo audio source separation based on time-frequency masking and multilevel thresholding," Digital Signal Processing, 2008, **18** (6): 960–976.

[9] 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, 2008, **16** (4): 748–756.

[10] M. Cobos and J. J. Lopez, "Two-microphone multispeaker localization based on a laplacian mixture model," Digital Signal Processing, 2011, **21**(1): 66–76.

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

[12] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," IEEE Transactions on Audio, Speech and Language Processing, 2006, **14** (4): 1462–1469.