AUDIO SOURCE SEPARATION WITH MAGNITUDE PRIORS: THE BEADS MODEL

Antoine Liutkus* *Christian Rohlfing*† *Antoine Deleforge*‡

*Inria and LIRMM, Montpellier, France
†Institut für Nachrichtentechnik, RWTH Aachen University, Germany
‡Inria Rennes - Bretagne Atlantique, France

ABSTRACT

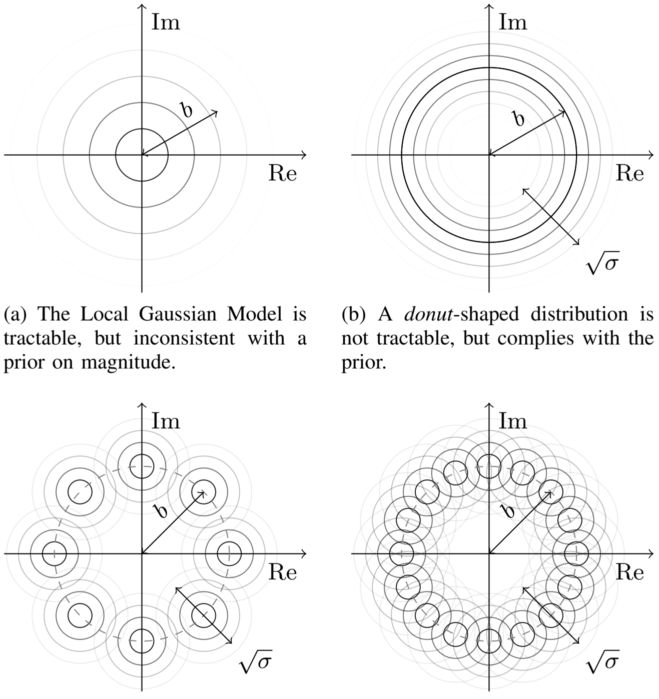
Audio source separation comes with the need to devise multichannel filters that can exploit priors about the target signals. In that context, experience shows that modeling magnitude spectra is effective. However, devising a probabilistic model on complex spectral data with a prior on magnitudes is non trivial, because it should both reflect the prior but also be tractable for easy inference. In this paper, we approximate the ideal donut-shaped distribution of a complex variable with approximately known magnitude as a Gaussian mixture model called BEADS (Bayesian Expansion Approximating the Donut Shape) and show that it permits straightforward inference and filtering while effectively constraining the magnitudes of the signals to comply with the prior. As a result, we demonstrate large improvements over the Gaussian baseline for multichannel audio coding when exploiting the BEADS model.

Index Terms— audio probabilistic model, magnitude, phase, source separation

I. INTRODUCTION

Audio source separation aims at processing an audio mixture x composed of I channels (e.g. stereo $I = 2$) so as to recover its J constitutive *sources* s_j [1]. It has many applications, including karaoke [2], upmixing [3], [4] and speech enhancement [5]. Usually, the sources are recovered by some time-varying *filtering* of the mixture, which is amenable to applying a $I \times I$ complex matrix $G(f,t)$ to each entry $x(f,t)$ of its Short-Term Fourier Transform (STFT) [6], [7], [3]. Devising good filters requires either dedicated models for sources power spectrograms [8], [9] or machine learning methods to directly predict effective filters [10].

The theoretical grounding for these linear filtering procedures boil down to considerations about second-order statistics for the sources STFT coefficients $s_j(f,t) \in \mathbb{C}$. When seen from a probabilistic perspective, this is often translated as picking a complex isotropic Gaussian distribution [11] for each $s_j(f,t) \in \mathbb{C}$, which is called the Local Gaussian Model (LGM) [12], [13] and is depicted on Figure 1a.



(a) The Local Gaussian Model is tractable, but inconsistent with a prior on magnitude.

(b) A *donut*-shaped distribution is not tractable, but complies with the prior.

(c) The BEADS model combines advantages of both ($C = 8$)...

(d) ...and approximates the *donut* shape well ($C = 16$).

Fig. 1: The BEADS probabilistic model for a complex variable with a magnitude close to b and squared uncertainty σ.

Although it enjoys tractability and easy inference, the main shortcoming of the LGM is that it gives highest probability to 0, which may appear as counterintuitive. Still, this comes as a consequence of the fact that stationarity makes all phases to be equally probable *a priori*, so that $\mathbb{E}[s_j(f,t)] = 0$. Combining the two first moments only, maximum entropy principles naturally lead us to pick the LGM [14].

However, this badly reflects the additional prior knowledge one often has on $s_j(f,t)$. The vast majority of methods use priors on its *magnitude*: $|s_j(f,t)|$ should be close to

This work was partly supported by the research programme KAMoulox (ANR-15-CE38-0003-01) funded by ANR, the French State agency for research.

some $b_j(f,t) > 0$ with squared uncertainty $\sigma_j(f,t)$. In that setting, the *donut*-shaped distribution shown in Figure 1b is much better than LGM because it gives its highest probability mass on the circle of radius $b_j(f,t)$. If $\sigma = 0$, we end up with the *phase unmixing problem* [15]. However, even with some uncertainty $\sigma > 0$, such a distribution suffers from non-tractability. In particular, it is not stable with respect to additivity, nor allows for simple posterior inference that would lead to straightforward filtering procedures.

In this paper, we translate our prior as a mixture of C identical Gaussian components evenly located over the circle of radius $b_j(f,t)$, yielding our proposed BEADS model (Bayesian Expansion Approximating the Donut Shape), depicted on Figure 1c. The most remarkable feature of BEADS is to allow both for straightforward filtering procedures while complying with priors on magnitude and possibly priors on phase. As such, it extends recent research on anisotropic modeling of complex spectra [16] by translating phase information into the choice of one particular component from the model. It thus appears as yet another way to incorporate phase information in source separation [17].

We illustrate the BEADS model in an informed source separation (ISS) setting, where the true sources are available at a first *coding* stage, that allows to compute good models to be used for separation at a *decoding* stage [18], [19], [20]. As demonstrated already before [21], this ISS setup is interesting because the separation parameters can be encoded very concisely, leading to effective instances of spatial audio object coding [22].

II. PROBABILISTIC MODEL

II-A. BEADS source model

The BEADS model is expressed as follows:

$$\begin{aligned}\mathbb{P}\left[s_j(f,t) \mid b_j(f,t), \sigma_j(f,t)\right] \\ = \sum_{c=1}^{C} \pi_j(c \mid f,t)\, \mathcal{N}\left(b_j(f,t)\,\omega^c, \sigma_j(f,t)\right), \quad (1)\end{aligned}$$

where $\mathcal{N}$ denotes here the complex isotropic Gaussian distribution[1], $\omega = \exp(i2\pi/C)$ is the C^{th} root of unity and $\pi_j(c \mid f,t)$ is the prior for the *phase* of source j at time-frequency (TF) bin (f,t): it indicates the probability that $s_j(f,t)$ is drawn from component c and hence that its phase is close to ω^c. While some phase unwrapping approach [23] may be used to set this prior, we take it as uniform here. The parameter $\sigma_j(f,t)$ stands for the *variance* of each component. It may be understood as the expected squared error of our estimate $b_j(f,t)$ for the magnitude. We can note that the LGM is equivalent to $C = 1$ and $b_j(f,t) = 0$, and that many beads tend to the donut shape, as shown in Figure 1d.

Now, we consider the joint prior distribution of the J independent sources. We need to consider all the C^J possible combinations for the components. Let $\mathbb{N}_C$ be the set of the first C natural numbers. We write $z(f,t) \in \mathbb{N}_C^J$ for the $J \times 1$ vector whose j^{th} entry $z_j(f,t) \in \mathbb{N}_C$ is the actual component drawn for source j. We define $\pi(\boldsymbol{c} \mid f,t)$ as the probability of each combination:

$$\forall \boldsymbol{c} \in \mathbb{N}_C^J, \pi(\boldsymbol{c} \mid f,t) \overset{\text{def}}{=} \mathbb{P}\left[z(f,t) = \boldsymbol{c}\right] = \prod_{j=1}^{J} \pi_j(c_j \mid f,t), \quad (2)$$

where c_j is the j^{th} entry of $\boldsymbol{c} \in \mathbb{N}_C^J$. We have $\sum_{\boldsymbol{c}} \pi(\boldsymbol{c} \mid f,t) = 1$. The joint prior distribution of the sources is given by:

$$\mathbb{P}\left[s(f,t) \mid \Theta\right] = \sum_{\boldsymbol{c} \in \mathbb{N}_C^J} \pi(\boldsymbol{c} \mid f,t)\, \mathcal{N}\left(\omega^{\boldsymbol{c}} \bullet b(f,t), [\sigma(f,t)]\right), \quad (3)$$

where $\omega^{\boldsymbol{c}}$ for $\boldsymbol{c} \in \mathbb{N}_C^J$ denotes the $J \times 1$ vector with entries ω^{c_j}, $a \bullet b$ denotes element-wise multiplication of the vectors a and b and $[v]$ denotes the diagonal matrix whose diagonal is the vector v. In words, the prior distribution for the sources under the BEADS model is a Gaussian Mixture Model (GMM), with weights $\pi(\boldsymbol{c} \mid f,t)$, which is reminiscent but different from the pioneering work in [24] that was limited to one source signal.

II-B. Mixture likelihood and separation

We take the mix as a convolutive mixture with the narrowband assumption, so that $x(f,t) \approx A(f)\, s(f,t)$, where $A(f)$ is the $I \times J$ mixing matrix at frequency bin f. Further exploiting the BEADS model (1), we get the marginal model of the mixture given all parameters Θ as:

$$\mathbb{P}\left[x(f,t) \mid \Theta\right] = \sum_{\boldsymbol{c} \in \mathbb{N}_C^J} \pi(\boldsymbol{c} \mid f,t)\, \mathcal{N}\left(\overline{x}_{\boldsymbol{c}}(f,t), \Sigma_x(f,t)\right), \quad (4)$$

with

$$\begin{cases} \overline{x}_{\boldsymbol{c}}(f,t) & = A(f)\left(\omega^{\boldsymbol{c}} \bullet b(f,t)\right) \\ \Sigma_x(f,t) & = A(f)\left[\sigma(f,t)\right] A(f)^{\star}. \end{cases} \quad (5)$$

Now, the real advantages of the BEADS model is that we can straightforwardly obtain the joint posterior distribution of the sources as:

$$\mathbb{P}\left[s(f,t) \mid x, \Theta\right] = \sum_{\boldsymbol{c} \in \mathbb{N}_C^J} \pi(\boldsymbol{c} \mid f,t,x)\, \mathcal{N}\left(\mu_{\boldsymbol{c}}, \Sigma_{\boldsymbol{c}}\right), \quad (6)$$

where the posterior statistics $\mu_{\boldsymbol{c}}$ and $\Sigma_{\boldsymbol{c}}$ for each combination $\boldsymbol{c}$ of the phases is:

$$\begin{cases} \mu_{\boldsymbol{c}}(f,t) = G(f,t)\left(x(f,t) - \overline{x}_{\boldsymbol{c}}(f,t)\right) + \omega^{\boldsymbol{c}} \bullet b(f,t) \\ \Sigma_{\boldsymbol{c}}(f,t) = \Sigma(f,t) = \left[\sigma(f,t)\right] - G(f,t)\, A(f)\left[\sigma(f,t)\right], \end{cases} \quad (7)$$

[1] The probability density function for the complex I-dimensional isotropic Gaussian is $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{\pi^I |\Sigma|} \exp\left(-(x-\mu)^{\star} \Sigma^{-1} (x-\mu)\right)$, where $|\Sigma|$ is the determinant of Σ [11].

Algorithm 1 BEADS decoder to update the phase configuration probabilities and perform separation.

```
Input: parameters $\Theta$, mixture $x(f,t)$, prior $\pi_j$, $C$.
Initialization: $\Sigma_{\tilde{\pi}} \leftarrow 0$, $\hat{s} \leftarrow 0$
For all $\boldsymbol{c}_n \in \mathbb{N}_C^J$:
    1) $\tilde{\pi}(\boldsymbol{c} \mid f,t,x) \leftarrow \pi(\boldsymbol{c} \mid f,t)\mathcal{N}(x(f,t) \mid \overline{x}_{\boldsymbol{c}}(f,t), \Sigma_x(f,t))$
    2) $\hat{s}(f,t) \mathrel{+}= \tilde{\pi}(\boldsymbol{c} \mid f,t,x)(\omega^{\boldsymbol{c}} \bullet b(f,t) - \overline{x}_{\boldsymbol{c}}(f,t))$
    3) $\Sigma_{\tilde{\pi}}(f,t) \mathrel{+}= \Sigma_{\tilde{\pi}}(f,t) + \tilde{\pi}(\boldsymbol{c} \mid f,t,x)$
Finalization:
    1) $\hat{s}(f,t) \leftarrow \frac{\hat{s}(f,t)}{\Sigma_{\tilde{\pi}}(f,t)}$
    2) $\hat{s}(f,t) \mathrel{+}= G(f,t)\,x(f,t)$
```

with the $J \times I$ Wiener gain $G(f,t)$ defined as:

$$G(f,t) = [\sigma(f,t)]\, A(f)^{\star} \left(A(f)[\sigma(f,t)]\, A(f)^{\star}\right)^{-1}.$$

The minimum mean squared error (MMSE) estimate for the sources thus becomes:

$$\begin{aligned}\mathbb{E}[s(f,t) \mid x, \Theta] &= G(f,t)\,x(f,t) \\ &+ \sum_{\boldsymbol{c} \in \mathbb{N}_C^J} \pi(\boldsymbol{c} \mid f,t,x)(\omega^{\boldsymbol{c}} \bullet b(f,t) - \overline{x}_{\boldsymbol{c}}(f,t)). \quad (8)\end{aligned}$$

In the case we know the true phases configuration $\boldsymbol{z}(f,t)$ or have estimated one at the decoder, this estimate (8) simplifies to $\hat{s}(f,t) = \mu_{\boldsymbol{z}(f,t)}$.

III. PARAMETERS ESTIMATION

In this section, we only consider the informed case, where the true source signals $s_j(f,t)$ and the mixing matrices $A(f)$ are known at the coder, while the mixture and the parameters $\Theta = \{b, \sigma, A\}$ are known at the decoder.

III-A. Decoder: posterior $\pi(\boldsymbol{c} \mid f,t,x)$ and separation

We show how the probabilities $\pi(\boldsymbol{c} \mid f,t)$ for the phase configurations may be updated to yield their posterior $\pi(\boldsymbol{c} \mid f,t,x)$ that fully exploit the BEADS model for constraining the phases. Dropping the (f,t) indices for readability, we have:

$$\forall \boldsymbol{c} \in \mathbb{N}_C^J, \pi(\boldsymbol{c} \mid x) = \frac{\pi(\boldsymbol{c})\mathcal{N}(x \mid \overline{x}_{\boldsymbol{c}}, \Sigma_x)}{\mathbb{P}[x]}. \quad (9)$$

This posterior probability may hence be expressed up to a normalizing constant independent of $\boldsymbol{c}$ as:

$$\pi(\boldsymbol{c} \mid x) \propto \tilde{\pi}(\boldsymbol{c} \mid x) = \pi(\boldsymbol{c}) \exp\left(-(x - \overline{x}_{\boldsymbol{c}})^{\star} \Sigma_x^{-1} (x - \overline{x}_{\boldsymbol{c}})\right), \quad (10)$$

which can straightforwardly be computed at the decoder with known parameters $\Theta = \{b, \sigma\}$. $\pi(\boldsymbol{c} \mid x)$ is obtained by normalization after computation for all $\boldsymbol{c} \in \mathbb{N}_J^C$.

Algorithm 1 summarizes the computations done at the decoder to perform estimation of the posterior probabilities for the phase configurations and separation.

III-B. Coder: amplitudes b and errors σ

The parameters to be learned at the coder are the amplitude priors $b_j(f,t)$ and the error models $\sigma_j(f,t)$. First, for saving bitrate and computing time, we only use BEADS for the F_0 frequency bands that have the highest energy in the mix, let them be $\mathcal{F}_0$, and simply pick $b_j(f,t) = 0$ for others. Then, for $f \in \mathcal{F}_0$ we compress them by picking a Nonnegative Tensor Factorization model [25]:

$$b_j(f,t) = \begin{cases} \sum_k W_b(f,k)\,H(t,k)\,Q_b(j,k) & \text{if } f \in \mathcal{F}_0 \\ 0 & \text{otherwise,} \end{cases} \quad (11)$$

in which case $\Theta_b = \{W_b, H, Q_b\}$ are small nonnegative $F_0 \times K$, $T \times K$ and $J \times K$, respectively. Same thing for σ, where we further reduce the number of parameters of the model by taking the same activations H in both cases, but this time, we model all frequencies.

In this section and for convenience, we will assume that the component $z_j(f,t)$ drawn for each source at each TF bin is known and equal to the one closest to $s_j(f,t)$. This simplification has the advantage of strongly reducing the computational cost of the estimation algorithm. Indeed, the BEADS model (1) then reduces to:

$$s_j(f,t) \mid z_j \sim \mathcal{N}\left(b_j(f,t)\,\omega^{z_j(f,t)}, \sigma_j(f,t)\right).$$

We can define the *relative* source $\epsilon_j(f,t)$:

$$\epsilon_j(f,t) \overset{\text{def}}{=} s_j(f,t)\left(\omega^{z_j(f,t)}\right)^{\star} \sim \mathcal{N}(b_j(f,t), \sigma_j(f,t)). \quad (12)$$

Provided $z_j(f,t)$ is correctly chosen as the component whose argument $\frac{2\pi z_j(f,t)}{C}$ is the closest to that of $s_j(f,t)$, we can furthermore safely assume that the real part $\mathcal{R}(\epsilon_j(f,t))$ of $\epsilon_j(f,t)$ is nonnegative.

Now, we detail the learning procedure we propose for b and σ. The strategy is to alternatively fix each of them and learn the other one.

Learning $b_j(f,t)$: Assume $\sigma_j(f,t)$ is kept fixed. The distribution (12) for the relative sources mean that we may estimate Θ_b using a weighted Euclidean method:

$$\Theta_b \leftarrow \underset{\Theta}{\text{argmin}} \sum_{f,t,j} \frac{|\mathcal{R}(\epsilon_j(f,t)) - b_j(f,t \mid \Theta)|^2}{\sigma_j(f,t)},$$

which can be done with a classical weighted NTF [26] scheme with the Euclidean cost function.

Learning $\sigma_j(f,t)$: If $b_j(f,t)$ is fixed we see from (12) that $\epsilon_j(f,t) - b_j(f,t)$ has an isotropic complex Gaussian distribution with variance $\sigma_j(f,t)$. This means Θ_σ can be estimated through:

$$\Theta_\sigma \leftarrow \underset{\Theta}{\text{argmin}} \sum_{f,t,j} d_{IS}\left(|\epsilon_j(f,t) - b_j(f,t)|^2 \parallel \sigma_j(f,t \mid \Theta)\right),$$

where $d_{IS}(a \parallel b)$ is the classical Itakura-Saito divergence for two nonnegative scalars a and b. This optimization is classical in the audio processing literature [27], [28], [29]. Considering that the activations H we take for σ are those for b, we only learn W_σ and Q_σ with it.

IV. EVALUATION

We evaluate the BEADS model through its performance for ISS, i.e. by displaying its average quality as a function of the bitrate required to transmit its parameters. To assess quality, we use BSSeval metrics [30]: SDR (Source to Distortion Ratio) and SIR (Source to Interference Ratio), both expressed in dB and higher for better separation. For normalization purpose, we compute δ-metrics, defined as the difference between the score and performance of oracle Wiener filtering, i.e. using true sources spectrograms [31].

The data consists of 10 excerpts of 30 s, taken from DSD100 database[2]. Each consists of $J = 4$ sources (vocals, bass, drums and accompaniment), sampled at 44.1 kHz. We generated either mono ($I = 1$) or stereo ($I = 2$) mixtures from these sources, through simple summation or anechoic mixing (delays+gains), respectively. STFT was conducted with 50 % overlap and a window size of 93 ms.
We evaluated the following methods:

- **BEADS oracle**: $\hat{s}(f,t) = \mu_{\boldsymbol{z}(f,t)}$.
- **BEADS point** using only the phase configuration $\hat{\boldsymbol{z}}$ that is most likely *a posteriori*.
- **BEADS** as given in Algorithm 1.
- **Itakura-Saito NTF** [32], with K components.

Given all these methods and data, our extensive evaluation consisted in trying the methods with $F_0 = 150$ frequency bands for BEADS magnitudes, $C = \{8, 16\}$ beads, and all methods were tried for $K \in [8, 128]$ NTF components. We picked 16 quantization levels for all parameters. Results were smoothed using LOESS [33] and are displayed on Figure 2.

An interesting fact we see on Figure 2 is that the oracle BEADS model significantly outperforms standard oracle Wiener filtering, even for very crude magnitude models $b_j(f,t)$. This can be seen by the fact that its δ-metrics get positive even at very small bitrates.

Then, we may notice that the δ-metrics appear as higher for mono than for stereo mixtures. In this respect, we should highlight that the absolute performance of oracle Wiener is of course higher for stereo (not shown on Figure 2), due to the knowledge of the mixing filters $A(f) \in \mathbb{C}$ that alone bring good separation already and actually some information about the phase of the sources. Adding additional spectral knowledge in that case is then less important than in the mono case, where it is crucial.

Now, we see a very clear improvement of BEADS as described in Algorithm 1 over classical NTF-ISS, of approximately 2 dB SDR and 5 dB SIR, at most bitrates. This significant boost in performance shows that BEADS helps a lot in predicting the source signals by adequately handling priors on magnitudes, which is the main result for this study.

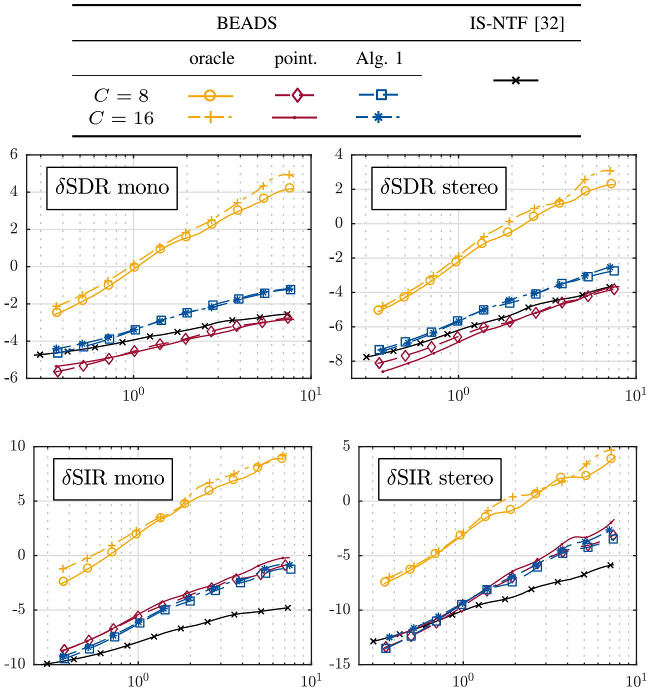


Fig. 2: BEADS for ISS on mono (left) or stereo (right) mixes. Metrics are δSDR (top) and δSIR (bottom). Units are kilobits/second/source (x-axis) and dB (y-axis).

Finally, Figure 2 also shows that the procedure for computing the phase posterior probabilities is not sufficient for correctly identifying the true phase configuration. This can be seen by the strong discrepancies between the BEADS point estimate and its oracle performance. While marginalization over the configuration as described in Algorithm 1 helps a lot in this respect, there is much room for improvement for parameter estimation of this model.

V. CONCLUSION

In this paper, we introduced BEADS as a convenient probabilistic model for complex data whose magnitude is approximately known. BEADS is a Gaussian Mixture Model where all components share the same variance and are scattered along a circle. While simple conceptually, BEADS comes with several advantages. First, it translates the delicate problem of modeling the phase into setting probabilities over a discrete set of components. Second, it allows for easy inference and, finally, it straightforwardly leads to effective filtering procedures. Although we demonstrated its potential in an audio-coding application, we believe it may also be useful in the blind separation setting when embedded in an Expectation-Maximization estimation procedure.

[2] http://sisec.inria.fr.

VI. REFERENCES

[1] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, pp. 107–115, May 2014.

[2] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 53–56, IEEE, 2012.

[3] C. Avendano and J.-M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Audio Engineering Society, 2002.

[4] A. Liutkus and P. Leveau, "Separation of music+ effects sound track from several international versions of the same movie," in *Audio Engineering Society Convention 128*, Audio Engineering Society, 2010.

[5] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge," *Computer Speech and Language*, vol. 46, no. C, pp. 605–626, 2017.

[6] E. Vincent, G. Jafari, A. Abdallah, D. Plumbley, and E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems* (W. Wang, ed.), pp. 162–185, IGI Global, 2010.

[7] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.

[8] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 550–563, March 2010.

[9] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118–1133, 2012.

[10] A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[11] R. Gallager, "Circularly Symmetric Complex Gaussian Random Vectors - A Tutorial," tech. rep., Massachusetts Institute of Technology, 2008.

[12] E. Vincent, S. Arberet, and R. Gribonval, "Underdetermined instantaneous audio source separation via local Gaussian modeling," vol. 5441/2009, (Paraty, Brésil), pp. pp 775–782, Springer-Verlag Berlin Heidelberg 2009, 2009.

[13] N. Duong, E. Vincent, and R. Gribonval, "Under-determined convolutive blind source separation using spatial covariance models," in *Acoustics, Speech and Signal Processing, IEEE Conference on (ICASSP'10)*, (Dallas, United States), pp. 9–12, Mar. 2010.

[14] E. T. Jaynes, *Probability theory: The logic of science*. Cambridge University Press, 2003.

[15] A. Deleforge and Y. Traonmilin, "Phase unmixing: Multichannel source separation with magnitude constraints," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 161–165, IEEE, 2017.

[16] P. Magron, R. Badeau, and B. David, "Phase-dependent anisotropic gaussian model for audio source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, (New Orleans, United States), Mar. 2017.

[17] N. Sturmel, L. Daudet, and L. Girin, "Phase-based informed source separation for active listening of music," in *15th International Conference on Digital Audio Effects (DAFx 2012)*, (York, United Kingdom), p. n/c, Sept. 2012.

[18] J. Nikunen, *Object-based Modeling of Audio for Coding and Source Separation*, vol. 1276. Tampere, Finland: Tampere University of Technology, 2015.

[19] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation : a comparative study," in *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[20] C. Rohlfing, J. Becker, and M. Wien, "NMF-based informed source separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 474–478, Mar. 2016.

[21] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-based informed source separation: Nonnegative tensor factorization approach," *IEEE Trans. on Audio, Speech and Language Processing*, 2012.

[22] J. Breebaart, J. Engdegård, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, J. Koppens, W. Oomen, B. Resch, E. Schuijers, *et al.*, "Spatial audio object coding (SAOC)-the upcoming MPEG standard on parametric object based audio coding," in *Audio Engineering Society Convention 124*, Audio Engineering Society, 2008.

[23] P. Magron, R. Badeau, and B. David, "Model-based STFT phase recovery for audio source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2018.

[24] S. Rennie, K. Achan, B. Frey, and P. Aarabi, "Variational speech separation of more sources than mixtures.," in *International Conference on Artificial Intelligence and Statistics, (AISTATS)*, 2005.

[25] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[26] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music.," in *SAPA@INTERSPEECH*, pp. 17–22, 2008.

[27] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Irish Signals and Systems Conference (ISSC)*, (Galway, Ireland), June 2008.

[28] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[29] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Computation*, vol. 23, pp. 2421–2456, Sep. 2011.

[30] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462 –1469, July 2006.

[31] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[32] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937 – 1949, 2012.

[33] W. Cleveland and S. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.