Bayesian anisotropic Gaussian model for audio source separation

Paul Magron, Tuomas Virtanen

BAYESIAN ANISOTROPIC GAUSSIAN MODEL FOR AUDIO SOURCE SEPARATION

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

In audio source separation applications, it is common to model the sources as circular-symmetric Gaussian random variables, which is equivalent to assuming that the phase of each source is uniformly distributed. In this paper, we introduce an anisotropic Gaussian source model in which both the magnitude and phase parameters are modeled as random variables. In such a model, it becomes possible to promote a phase value that originates from a signal model and to adjust the relative importance of this underlying model-based phase constraint. We conduct Bayesian inference of the model through the derivation of an expectation-maximization algorithm for estimating the parameters. Experiments conducted on realistic music songs for a monaural source separation task, in an scenario where the variance parameters are assumed known, show that the proposed approach outperforms state-of-the-art techniques.

Index Terms— Anisotropic Gaussian model, Bayesian inference, expectation-maximization, audio source separation

1. INTRODUCTION

The goal of audio source separation [1] is to extract underlying *sources* that add up to form an observable audio *mixture*. To address this issue, many techniques act on a time-frequency (TF) representation of the data, such as the short-term Fourier transform (STFT), since the structure of audio signals is more prominent in that domain.

A popular approach to tackle this problem is to frame it in a probabilistic framework, where the sources are modeled as random variables [2]. Those variables' parameters are further structured by means of a model, such as nonnegative matrix factorization (NMF) [3], kernel additive models [4] or deep neural networks (DNNs) [5]. Most approaches consider circular-symmetric (or *isotropic*) Gaussian distributions [3], which is equivalent to assuming that the phase of each source is uniformly distributed. In such a framework, the sources are typically estimated in a minimum mean square error (MMSE) sense by means of a Wiener filter, which assigns the phase of the original mixture to each extracted source. However, even if this filter yields quite satisfactory sounding estimates in practice [3, 6], it has been pointed out [7] that when sources overlap in the TF domain, it is responsible for residual interference and artifacts in the separated signals.

Indeed, even if the phase may globally appear as uniform [8], it holds some underlying structure that can be exploited. For instance, the model of mixtures of sinusoids leads to explicit constraints between the phases of adjacent TF bins [9]. Such an approach has been exploited in speech enhancement [10], audio restoration [9] and source separation [11, 12]. Drawing on those observations, we proposed in a preliminary work [13] to model the sources with anisotropic Gaussian (AG) variables, i.e., where the phase is no longer uniform. In such a model, one can promote a phase value which is obtained by exploiting the sinusoidal model. MMSE estimation results in an anistropic Wiener (AW) filter, which optimally combines the mixture phase and the underlying phase model.

In this paper, we introduce an AG model that differs from [13] in two ways. Firstly, we model the magnitudes as random variables instead of deterministic parameters that were estimated beforehand in [13]. This is a suitable choice when the magnitudes must be estimated along with the phases. Secondly, we model the phase location parameter as a random variable with a Markov chain prior structure, instead of unwrapping it over time frames in a deterministic fashion [13]. This allows us to adjust the relative importance of the underlying phase constraint. In such a model, it becomes possible to perform Bayesian inference of the parameters thanks to an expectation-maximization (EM) algorithm. Experiments conducted on a source separation task show that this approach outperforms our preliminary AW technique [13] and the consistent Wiener filter [14], which reaches improved phase recovery by exploiting the redundancy of the STFT instead of a model-based phase constraint.

This paper is organized as follows. Section 2 presents the AG model. Section 3 introduces the EM procedure for estimating the model parameters. Section 4 experimentally validates the potential of this method for an audio source separation task. Finally, Section 5 draws some concluding remarks.

2. ANISOTROPIC GAUSSIAN MODEL

Let $\mathbf{X} \in \mathbb{C}^{F \times T}$ be the STFT of a single-channel audio signal, where F and T are the numbers of frequency channels and time frames. $\mathbf{X}$ is the linear and instantaneous mixture of J sources $\mathbf{S}_j \in \mathbb{C}^{F \times T}$, such that for all TF bins ft,

$$x_{ft} = \sum_{j=1}^{J} s_{j,ft}. \tag{1}$$

Since all TF bins are treated similarly, we remove the indices ft when appropriate for more clarity.

2.1. Anisotropic Gaussian sources

We assume that each source s_j follows a complex normal distribution: $s_j \sim \mathcal{N}(m_j, \Gamma_j)$, where m_j is the mean of s_j and:

$$\Gamma_j = \begin{pmatrix} \gamma_j & c_j \\ \bar{c}_j & \gamma_j \end{pmatrix} \tag{2}$$

is its covariance matrix, where γ_j and c_j are the variance and relation term of s_j, and $\bar{z}$ denotes the complex conjugate of z. Many previous studies model the sources as circular-symmetric (or *isotropic*)

The work of P. Magron was partly supported by the Academy of Finland, project no. 290190.

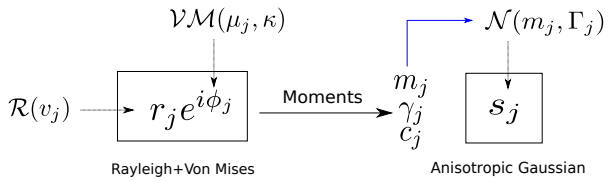

Fig. 1. Design of the AG model. We first model the magnitudes and phases as Rayleigh and Von Mises random variables. The moments in this model are then used to define the equivalent AG model.

variables [3, 14] (i.e., such that $m_j = c_j = 0$), which is equivalent to assuming that the phase of each source is uniformly distributed. Drawing on a preliminary work [13], we rather consider here that the phases ϕ_j should be distributed around some favored value μ_j with a concentration parameter $\kappa \in [0, +\infty[$. We therefore model the phases as Von Mises [15] random variables with location parameter μ_j and concentration parameter κ.

In [13], we assumed that the magnitudes were deterministic parameters estimated beforehand. Here, we propose to model the magnitudes r_j as Rayleigh random variables (which corresponds to the modulus of an isotropic complex normal distribution), with a dispersion parameter v_j (which is an estimate of the source power). This results in a Rayleigh + von Mises (RvM) model, which however is not tractable (the density of the mixture does not admit a closed-form expression). Therefore, following the methodology of [13], we compute the moments in the RvM model which are then used to define an equivalent Gaussian model, as illustrated in Fig. 1:

$$m_j = \lambda\sqrt{v_j}e^{i\mu_j}, \gamma_j = (1-\lambda^2)v_j, \text{ and } c_j = \rho v_j e^{i2\mu_j}, \quad (3)$$

with:

$$\lambda = \frac{\sqrt{\pi}}{2}\frac{I_1(\kappa)}{I_0(\kappa)} \text{ and } \rho = \frac{I_2(\kappa)}{I_0(\kappa)} - \lambda^2, \quad (4)$$

where I_n is the modified Bessel function of the first kind of order n. In particular, if $\kappa = 0$, then $\lambda = \rho = 0$, and consequently $m_j = c_j = 0$: the distributions become isotropic. Otherwise, it holds a property of *anisotropy*, hence the name of the model. Finally, $x \sim \mathcal{N}(m_x, \Gamma_x)$ with $m_x = \sum_{j=1}^{J} m_j$ and $\Gamma_x = \sum_{j=1}^{J} \Gamma_j$.

2.2. Phase model

We propose to incorporate a priori phase information on μ_j from a sinusoidal model, which is widely used for representing audio signals [10, 11]. Each source in the time domain is modeled as a sum of sinusoids. Assuming there is at most one sinusoid per frequency channel, let us denote by $\nu_{j,ft}$ the normalized frequency in channel f. It can be shown [9] that the phase μ_j follows the unwrapping equation in the TF domain:

$$\mu_{j,ft} \approx \mu_{j,ft-1} + 2\pi l \nu_{j,ft}, \quad (5)$$

where l is the hop size of the STFT. As in [9], we estimate the frequencies $\nu_{j,ft}$ by means of a quadratic interpolated FFT [16] on the log-spectra of the sources at each time frame, in order to account for slow variations of the frequencies.

We propose to enforce this property by means of a Markov chain prior structure, as done in [17] to enforce the smoothness of the activation matrix in an NMF model. We have, for each source:

$$p(\mu_j) = \prod_{f=0}^{F-1} p(\mu_{j,f0}) \prod_{t=1}^{T-1} p(\mu_{j,ft}|\mu_{j,ft-1}). \quad (6)$$

We then propose the following choice, for $t \neq 0$:

$$\mu_{j,ft}|\mu_{j,ft-1} \sim \mathcal{VM}(\mu_{j,ft-1} + 2\pi l \nu_{j,ft}, \tau), \quad (7)$$

where $\mathcal{VM}$ denotes the Von Mises distribution. In this way, the phase location parameter approximately follows the sinusoidal model (5). The initial distribution in each frequency channel $p(\mu_{j,f0})$ is Jeffrey's non-informative prior. Therefore:

$$\log(p(\mu)) \stackrel{c}{=} \tau \sum_{j,f,t} \Re\left(e^{i\mu_{j,ft}} e^{-i\mu_{j,ft-1} - 2i\pi l \nu_{j,ft}}\right), \quad (8)$$

where $\stackrel{c}{=}$ denotes equality up to an additive constant and $\Re$ is the real part.

3. BAYESIAN INFERENCE

We estimate the model parameters $\Theta = \{\{v_j\}_j, \{\mu_j\}_j\}$ in a maximum a posteriori sense, which consists in maximizing the log-posterior distribution:

$$\mathcal{C}_{\text{MAP}}(\Theta) = \log p(\mathbf{X}|\Theta) + \log p(\Theta), \quad (9)$$

where $p(\mathbf{X}|\Theta)$ is the likelihood of the data and $p(\Theta)$ the priors on the parameters. In this work, we only exploit some prior information about the phase, therefore $\log p(\Theta)$ is given by (8).

3.1. EM framework

Since the direct maximization of the criterion (9) is more involved that in classical isotropic models [3], we propose to adopt an EM [18] strategy which consists in maximizing a lower bound of the log-posterior distribution, given by:

$$Q^{\text{MAP}}(\Theta, \Theta') = Q^{\text{ML}}(\Theta, \Theta') + \log p(\Theta), \quad (10)$$

where Θ' contains the current set of estimated parameters and Q^{ML} is the conditional expectation of the complete-data log-likelihood:

$$Q^{\text{ML}}(\Theta, \Theta') = \int p(\mathbf{Z}|\mathbf{X}; \Theta') \log p(\mathbf{X}, \mathbf{Z}; \Theta) d\mathbf{Z}, \quad (11)$$

where $\mathbf{Z}$ denotes a set of latent (hidden) variables. Due to the mixing constraint (1), we consider, as in [14, 19], a reduced set of $J' = J - 1$ free variables $\mathbf{Z} = \mathbf{S} = \{\mathbf{s}_{ft}\}_{ft}$, where we note $\mathbf{s}_{ft} = [s_{1,ft}, ..., s_{J',ft}]^T$ and where ^{T} denotes matrix transposition.

The EM algorithm consists in alternatively computing the functional Q^{MAP} given the current set of parameters Θ' (E-step) and maximizing it with respect to Θ (M-step). This is proven [18] to increase the value of the criterion (9).

3.2. E-step

Since all $\{s_{j,ft}\}_{j=1}^{J'}$ are independent Gaussian variables, $\mathbf{s}_{ft}$ is a Gaussian vector $\mathbf{s}_{ft} \sim \mathcal{N}(\mathbf{m}_{ft}, \mathbf{\Sigma}_{z,ft})$ with $\mathbf{m}_{ft} = [m_{1,ft}, ..., m_{J',ft}]^T$ and $\mathbf{\Sigma}_{s,ft} = \text{diag}([\Gamma_{1,ft}, ..., \Gamma_{J',ft}])$ is a block-diagonal matrix.

It can be shown [20] that $\mathbf{S}|\mathbf{X}$ follows a multivariate complex normal distribution $\mathcal{N}(\mathbf{m}'_{ft}, \mathbf{\Xi}_{ft})$. The posterior mean vector $\mathbf{m}'_{ft} = [m'_{1,ft}, ..., m'_{J',ft}]^T$ is given by [13]:

$$\underline{m}'_{j,ft} = \underline{m}_{j,ft} + \Gamma_{j,ft}\Gamma^{-1}_{x,ft}(\underline{x}_{ft} - \underline{m}_{x,ft}), \quad (12)$$

where $\underline{x} = \begin{pmatrix} x & \bar{x} \end{pmatrix}^T$. The diagonal blocks in the posterior covariance matrix $\mathbf{\Xi}_{ft}$ provide the posterior covariance for each source:

$$\Gamma'_{j,ft} = \Gamma_{j,ft} - \Gamma_{j,ft}\Gamma^{-1}_{x,ft}\Gamma_{j,ft}. \quad (13)$$

Thanks to (12) and (13), we can compute the posterior mean, variance and relation term of the sources respectively denoted m'_j, γ'_j and c'_j. Due to the lack of space, we cannot detail here the full computation of (11) (it will be provided in a future study): in a nutshell, it consists in using the same algebra as in [21], which leads to:

$$\begin{aligned} Q^{\text{ML}}(\Theta,\Theta') \stackrel{c}{=} &-\sum_{j=1}^{J}\sum_{f,t}\log(\sqrt{|\Gamma_{j,ft}|}) \\ &+\frac{1}{|\Gamma_{j,ft}|}\left(\gamma_{j,ft}(|m'_{j,ft}-m_{j,ft}|^2+\gamma'_{j,ft})\right) \\ &-\frac{1}{|\Gamma_{j,ft}|}\left(\Re(\bar{c}_{j,ft}((m'_{j,ft}-m_{j,ft})^2+c'_{j,ft}))\right), \end{aligned} \quad (14)$$

where $|\Gamma_{j,ft}|$ is the determinant of $\Gamma_{j,ft}$.

3.3. M-step: variance parameters

Let us first estimate the variance parameters. Since Q^{MAP} is equal to Q^{ML} up to the log-prior on the phase, which does not depend on the variance parameters, we have, from (10) and (14):

$$Q^{\text{MAP}}(\Theta|\Theta') \stackrel{c}{=} -\sum_{j=1}^{J}\sum_{f,t}\log(v_{j,ft})+\frac{p_{j,ft}}{v_{j,ft}}+\frac{q_{j,ft}}{\sqrt{v_{j,ft}}}, \quad (15)$$

with:

$$p=\frac{(1-\lambda^2)\left(\gamma'+|m'|^2\right)-\rho\Re\left(e^{-2i\mu}(c'+m'^2)\right)}{(1-\lambda^2)^2-\rho^2}, \quad (16)$$

and:

$$q=\frac{2\lambda(\rho-1+\lambda^2)}{(1-\lambda^2)^2-\rho^2}\Re\left(e^{-i\mu}m'\right), \quad (17)$$

where we removed the indices j,ft for brevity[1]. The derivative of Q^{MAP} with respect to v is:

$$\nabla_v Q^{\text{MAP}}(\Theta|\Theta')=-\frac{1}{v}+\frac{p}{v^2}+\frac{1}{2}\frac{q}{v\sqrt{v}}. \quad (18)$$

Setting this derivative to 0 and multiplying by v^2 leads to a second-order polynomial equation in the variable $\sqrt{v}$. The only positive root then provides us the update on v:

$$v=\frac{1}{16}\left(q+\sqrt{16p+q}\right)^2. \quad (19)$$

Remark: Here, the variance parameters v_j are supposed to be unconstrained. In more realistic applications, it becomes necessary to constrain it by means of an appropriate fitting model (e.g., NMF [3] or DNNs [5]) in order to yield good quality results.

3.4. M-step: phase parameters

Let us now derive the updates on the phase parameters. We rewrite the functional (14) by removing the terms that do not depend on the phase parameters, which leads to:

$$Q^{\text{ML}}(\Theta|\Theta') \stackrel{c}{=} \sum_{j=1}^{J}\sum_{f,t}\Re\left(\alpha_{j,ft}e^{-2i\mu_{j,ft}}+\beta_{j,ft}e^{-i\mu_{j,ft}}\right), \quad (20)$$

with:

$$\alpha_{j,ft}=\frac{\rho}{((1-\lambda^2)^2-\rho^2)v_{j,ft}}(c'_{j,ft}+m'^2_{j,ft}), \quad (21)$$

and:

$$\beta_{j,ft}=\frac{2\lambda(1-\lambda^2-\rho)}{((1-\lambda^2)^2-\rho^2)\sqrt{v_{j,ft}}}m'_{j,ft}. \quad (22)$$

Therefore, adding the log-prior over the phase parameters (8) leads to independently maximizing the following functionals:

$$g_{j,ft}(\mu_{j,ft})=\Re\left(\alpha_{j,ft}e^{-2i\mu_{j,ft}}+\tilde{\beta}_{j,ft}e^{-i\mu_{j,ft}}\right), \quad (23)$$

with respect to $\mu_{j,ft}$, and where:

$$\tilde{\beta}_{j,ft}=\beta_{j,ft}+\tau\left(e^{i\mu_{j,ft-1}+2i\pi l\nu_{j,ft}}+e^{i\mu_{j,ft+1}-2i\pi l\nu_{j,ft+1}}\right). \quad (24)$$

Let us remove the indexes j,ft in what follows for more clarity. It can be shown that maximizing g involves solving a fourth-order polynomial in the variable $e^{i\mu}$. Tractable solutions exist, but they are not straightforward to implement an it requires further operations to determine which root maximizes g, leading to a quite computationally intensive procedure. Instead, since we experimentally observed that $|\alpha| << |\beta|$, we propose to approximate (23) by:

$$\tilde{g}(\mu)=\Re\left(\tilde{\beta}e^{-i\mu}\right)=|\tilde{\beta}|\cos(\mu-\angle\tilde{\beta}), \quad (25)$$

which is easily maximized[2] for $\mu=\angle\tilde{\beta}$. Note that this update depends on the values of the phase parameter in frames $t-1$ and $t+1$, so it has to be applied sequentially over time frames (which is common when using Markov chain priors such as in [17]).

3.5. Full procedure

The full EM procedure is summarized in Algorithm 1. One final E-step is performed after looping in order to estimate the sources with the most up-to-date set of parameters. Note that if the initial variance estimate is reliable enough, one can choose to only update the phase parameters (i.e., one skips lines 14 and 16 in Algorithm 1).

4. EXPERIMENTAL RESULTS

In this section, we experimentally asses the potential of the proposed model for a monaural audio source separation task.

4.1. Setup

We consider 100 music song excerpts from the DSD100 database, a semi-professionally mixed set of music songs used for the SiSEC 2016 campaign [22]. Each excerpt is 10 seconds long and is made up of $J=4$ sources: `bass`, `drum`, `vocals` and `other` (which may contain various instruments such as guitar, piano...). The database consists of two subsets of 50 songs (learning and test sets).

The signals are sampled at $F_s=44100$ Hz and the STFT is computed with a 92 ms long Hann window and 75 % overlap.

In these experiments, we only inquire about the potential of adding some phase information within a probabilistic model. Therefore, the variance parameters v_j are assumed known (they are equal to the ground truth power spectrograms of sources) and are not updated in Algorithm 1. The frequencies ν_j are computed as detailed

[1] Quite interestingly, in the isotropic case (i.e., when $\kappa=0$), we see from (4) that $\lambda=\rho=0$, and therefore $q_j=0$ and $p_j=\gamma'_j+|m'_j|^2$, which is the posterior power of s_j. Then, we recognize in (15) the Itakura-Saito divergence between p_j and v_j, as in [3].

[2] Since this update no longer maximizes (23), but instead increases it, we should actually refer to the procedure as a generalized EM algorithm.

Algorithm 1: EM algorithm for AG model estimation

Inputs: Mixture $\mathbf{X} \in \mathbb{C}^{F \times T}$,
Phase parameters κ and $\tau \in \mathbb{R}_+$,
Normalized frequencies $\nu \in \mathbb{R}^{J \times F \times T}$.
Initialization: compute λ and ρ with (4),
Compute m, γ and c with (3).
while stopping criterion not reached **do**
  % E-step
  $(m_x, \gamma_x, c_x) = \sum_{j=1}^{J} (m_j, \gamma_j, c_j)$,
  Update m' with (12),
  Update γ' and c' with (13),
  Update p with (16) and q with (17).
  % M-step
  Update v with (19),
  Update β with (22).
  for $t = 1$ to $T - 2$ **do**
    $\forall (j, f)$, update $\tilde{\beta}_{j,ft}$ with (24),
    $\mu_{j,ft} = \angle \tilde{\beta}_{j,ft}$.
  end
  Update m, γ and c with (3).
end
$(m_x, \gamma_x, c_x) = \sum_{j=1}^{J} (m_j, \gamma_j, c_j)$,
Update m' with (12),
Outputs: $m' \in \mathbb{C}^{J \times F \times T}$.

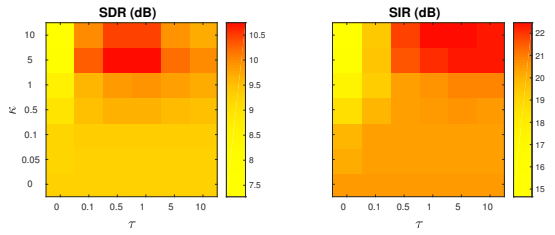

Fig. 2. Influence of the phase parameters κ and τ on the source separation quality (the SAR is not shown since it exhibits a similar behavior to SDR).

in Section 2.2 and provided as inputs in Algorithm 1. The stopping criterion for the algorithm is set to 40 iterations (performance is not further improved beyond).

Source separation quality is measured with the signal-to-distortion, signal-to-interference, and signal-to-artifact ratios (SDR, SIR, and SAR) [23] expressed in dB, where only a rescaling (not a refiltering) of the reference is allowed.

4.2. Impact of the phase parameters

As in [13], we first study the impact of the phase parameters κ and τ on the separation quality. Let us consider the 50 songs from the learning set and run the proposed procedure for various values of those parameters. The averaged results are presented in Fig. 2.

Those results show that for non-null values of the phase parameters, the proposed approach can outperform a phase-unaware approach (for which $\kappa = \tau = 0$) according to the SDR, SIR and SAR. However, a compromise between those criteria must be reached: indeed, the best SIR and the best SDR/SAR are not obtained for the same values of the phase parameters. This observation is reminiscent of some previous works, such as [12]: it is generally observed that promoting the sinusoidal model-based phase constraint leads to reducing interference between sources, but at the cost of some artifacts. Therefore, we choose the values $\kappa = 5$ and $\tau = 0.5$ for the next experiment, since it seems to be a good compromise in terms of overall separation quality.

	Wiener	CW	AW	Proposed
SDR	8.6	**10.5**	9.7	10.0
SIR	19.1	**22.2**	21.7	20.2
SAR	9.1	**11.0**	10.0	10.5
OPS	19.2	19.7	23.0	**23.3**
TPS	28.4	30.4	32.9	**32.9**
IPS	34.7	34.5	37.7	**38.9**
APS	30.6	31.0	**34.8**	34.1

Table 1. Average source separation performance.

4.3. Comparison with other methods

As comparison references, we test Wiener filtering [3] and consistent Wiener (CW) filtering [14]. We also test the AW filtering from our previous work [13]. Note that Wiener filtering corresponds to our approach in the isotropic case, i.e., when $\kappa = 0$ and therefore $m_j = c_j = 0$. CW and AW depend on a parameter which either promotes consistency or anisotropy, and which is learned as in Section 4.2. The results averaged over the test dataset are presented in Table 1.

From the first three lines of Table 1, it can be seen that the proposed method globally outperforms Wiener filtering, and also performs better than AW in terms of SDR and SAR. However, it performs worse than CW according to those three indicators. Nonetheless, an informal perceptual evaluation shows that our method may yield better results in terms of perceptual quality (the interested reader can listen at the sounds excerpts available at [24]). In particular, the `bass` track is neater when estimated with our method compared to the others, and the `drum` track contains less artifacts.

Therefore, we also computed the PEASS score [25], which provides a novel set of criteria that is built upon a subjective evaluation of source separation quality, and designed to better match perception than the SDR, SIR and SAR. The resulting criteria are the overall, target-related, interference-related and artifacts-related Perceptual Scores (OPS, TPS, IPS and APS). The corresponding results are presented in the four last lines of Table 1. According to those, the proposed approach outperforms CW, and yields results similar to or better than AW (except in terms of APS). These results are consistent with our perceptual evaluation, and show the potential of the proposed AG model for a phase-aware audio source separation task.

5. CONCLUSION

In this paper, we introduced a Bayesian framework to estimate latent variables in mixtures of anisotropic Gaussian sources. Such a model permits us to exploit some model-based prior information about the phase, and outperforms our preliminary model [13] and the state-of-the art consistent Wiener filtering [14]. Therefore, this is a novel step towards a complete phase-aware separation system.

In future work, we will focus on the estimation of the variance parameters v_j, since they were assumed known in the evaluation of this paper. In realistic scenarios, it becomes necessary to structure it by means of a variance fitting model. Examples of such models are DNNs, as already used in a multichannel framework with isotropic Gaussian variables [5] or NMF which has shown promising results for supervised separation tasks [26].

6. REFERENCES

[1] P. Comon and C. Jutten, *Handbook of blind source separation: independent component analysis and applications*. Academic press, 2010.

[2] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2005.

[3] 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, no. 3, pp. 793–830, March 2009.

[4] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel additive models for source separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4298–4310, August 2014.

[5] A. 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, September 2016.

[6] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[7] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[8] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007.

[9] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *Proc. of European Signal Processing Conference (EUSIPCO)*, August 2015.

[10] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, December 2014.

[11] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014.

[12] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: application to audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016.

[13] ——, "Phase-dependent anisotropic Gaussian model for audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[14] J. Le Roux and E. Vincent, "Consistent Wiener filtering for audio source separation," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 217–220, March 2013.

[15] K. V. Mardia and P. J. Zemroch, "Algorithm AS 86: The Von Mises distribution function," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 2, pp. 268–272, 1975.

[16] M. Abe and J. O. Smith, "Design criteria for simple sinusoidal parameter estimation based on quadratic interpolation of FFT magnitude peaks," in *Audio Engineering Society Convention 117*, May 2004.

[17] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 538–549, March 2010.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[19] P. Magron, J. Le Roux, and T. Virtanen, "Consistent anisotropic Wiener filtering for audio source separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017.

[20] B. Picinbono, "Second-order complex random vectors and normal distributions," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2637–2640, October 1996.

[21] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, pp. 2453–2465, 2016.

[22] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. of International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, 2017.

[23] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[24] http://www.cs.tut.fi/~magron/demos/demo_ICASSP2018.html.

[25] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, September 2011.

[26] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of ICA*, September 2007.