PHASE-DEPENDENT ANISOTROPIC GAUSSIAN MODEL FOR AUDIO SOURCE SEPARATION

Paul Magron *Roland Badeau* *Bertrand David*

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013, Paris, France
<firstname>.<lastname>@telecom-paristech.fr

ABSTRACT

Phase reconstruction of complex components in the time-frequency domain is a challenging but necessary task for audio source separation. While traditional approaches do not exploit phase constraints that originate from signal modeling, some prior information about the phase can be obtained from sinusoidal modeling. In this paper, we introduce a probabilistic mixture model which allows us to incorporate such phase priors within a source separation framework. While the magnitudes are estimated beforehand, the phases are modeled by Von Mises random variables whose location parameters are the phase priors. We then approximate this non-tractable model by an anisotropic Gaussian model, in which the phase dependencies are preserved. This enables us to derive an MMSE estimator of the sources which optimally combines Wiener filtering and prior phase estimates. Experimental results highlight the potential of incorporating phase priors into mixture models for separating overlapping components in complex audio mixtures.

Index Terms— Phase reconstruction, Von Mises distribution, anisotropic Gaussian model, phase unwrapping, source separation

1. INTRODUCTION

Source separation consists in extracting underlying components called *sources* that add up to form an observable signal called *mixture*. A variety of audio source separation techniques acts in the Time-Frequency (TF) domain, exploiting the particular structure of music signals. For instance, the family of techniques based on Nonnegative Matrix Factorization (NMF) [1] is often applied to spectrogram-like representations, such as the modulus of the Short-Time Fourier Transform (STFT). It has proved to provide a promising framework for audio source separation [2, 3].

However, when it comes to resynthesizing time signals, obtaining the phase of the corresponding complex-valued STFT is necessary, and is still an open issue [4,5]. In the single-channel source separation framework, a common practice consists in applying Wiener-like filtering [3]: the phase of the mixture is given to each extracted component. Alternatively, a consistency-based approach can be used for phase recovery [6]. That is, a complex-valued matrix is iteratively computed in order to maximize its consistency, i.e. to bring it as close as possible to the STFT of a time signal. It has however been pointed out [7] that consistency-based approaches provide poor results in terms of audio quality. Besides, Wiener filtering fails to provide good results when sources overlap in the TF domain. There were some attempts [8–11] to overcome the limitations of those two approaches by combining them in a unified framework. Consistent Wiener filtering [11] has shown to be the most promising candidate for this task. Alternatively, phase reconstruction from spectrograms can be performed using phase models based on signal analysis. For instance, the widely used model of mixtures of sinusoids [12] can lead to explicit constraints for phase reconstruction that exploit the relationships between adjacent TF bins [13]. Such an approach has been exploited in the phase vocoder algorithm [14], speech signal reconstruction [15, 16], audio restoration [13] and integrated into several Complex NMF frameworks [17, 18] for audio source separation.

However, using such a phase unwrapping prior to estimate the components without accounting for the mixture phase may lead to audible artifacts in the reconstructed signals [13]. It is then necessary to design a model which accounts for both the mixture phase and the phase prior. Such mixture models have been proposed in the literature [19, 20], but they are generally restricted to mixtures of 2 sources (speech and noise) in a speech enhancement framework. In this paper, we define a probabilistic mixture model where the phases are modeled by Von Mises random variables, a circular distribution that allows us to incorporate some prior information about the phases. Since in this model the Probability Density Function (PDF) of the mixture is not tractable, we propose to approximate it by an anisotropic Gaussian model whose moments are the same ones as in the Von Mises model. This new model benefits from being phase-dependent and fully tractable. We further derive an estimator of the sources which is optimal in a Minimum Mean Square Error (MMSE) sense. Experiments on realistic music songs show that this approach delivers results that are similar to those provided by the consistent Wiener filtering technique in terms of source separation quality, with a significantly lower computational cost.

This paper is organized as follows. Section 2 presents the mixture model based on Von Mises phase priors. It is then approximated by the anisotropic Gaussian model in Section 3, where an MMSE estimator of the sources is obtained. Section 4 experimentally validates the potential of this method for an audio source separation task. Finally, Section 5 draws some concluding remarks.

2. VON MISES MIXTURE MODEL

Let $X \in \mathbb{C}^{F \times T}$ be the STFT of an audio signal. X is the mixture of K sources Z_k, such that $X = \sum_k Z_k$. The problem of source separation consists in obtaining an estimator of the sources Z_k. Assuming that a prior estimate of the magnitudes V_k is available (e.g. after a preliminary NMF [1]), one only needs to estimate their phases $\phi_k = \angle Z_k$, where $\angle(.)$ denotes the complex argument. Since some prior information about these phases can be obtained [13], we propose to incorporate it in a probabilistic model. Given that all TF bins are treated independently, we consider a bin indexed by (f, t) and we remove the indexes in what follows for a clarity purpose.

This work is partly supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR-13-CORD-0008-02).

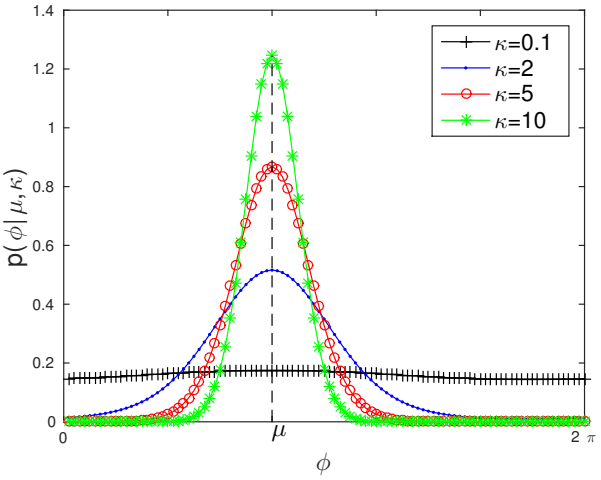

Fig. 1. Von Mises PDF for a location parameter μ and several values of the concentration parameter κ.

2.1. Von Mises phase

The most commonly used distribution in circular statistics is the Von Mises distribution, since it is the maximum entropy distribution for circular data and a good approximation of the wrapped normal distribution with a tractable PDF [21]. The wrapped normal distribution has notably been used in [22] for speech modeling, however its PDF does not have a simple closed-form expression. The Von Mises distribution, denoted $\mathcal{VM}(\mu, \kappa)$, is parametrized by a location parameter $\mu \in [0; 2\pi[$ and a concentration parameter $\kappa \in]0; +\infty[$. Its PDF for the angle ϕ is given by:

$$p(\phi|\mu, \kappa) = \frac{e^{\kappa \cos(\phi - \mu)}}{2\pi I_0(\kappa)}, \tag{1}$$

where I_n is the modified Bessel function of the first kind of order n [23]. It is illustrated in Fig. 1. In particular, if $\kappa \to 0$, the Von Mises distribution becomes equivalent to the uniform distribution. Contrarily, if $\kappa \to +\infty$, it becomes equivalent to a Dirac delta function centered at μ. The proposed mixture model is then:

$$X = \sum_k Z_k \text{ with } Z_k = V_k e^{i\phi_k} \text{ and } \phi_k \sim \mathcal{VM}(\mu_k, \kappa_k), \tag{2}$$

where μ_k is a phase prior that can be obtained e.g. by the phase unwrapping algorithm [13], and κ_k is a weight which promotes this prior. Similar models have been proposed, notably in [19, 20]. However, they were restricted to mixtures of $K = 2$ sources in a speech enhancement framework. Nevertheless, those recent papers suggest that the Von Mises distribution is an appropriate tool for modeling the STFT phase of audio signals.

2.2. Phase unwrapping prior

The sinusoidal model is an ubiquitous mixture model for representing audio signals [12]. Let us consider a source indexed by $k \in [\![1; K]\!]$ which is modeled as a sum of sinusoids. Assuming there is at most one sinusoid per frequency channel, let us denote by $\nu_k(f, t)$ the normalized frequency in channel f and time frame t. Ω_k denotes the onset frame indexes for source k. It can be shown [13] that the phase μ_k of the k-th source follows the unwrapping equation: $\forall f \in [\![0; F-1]\!]$ and $\forall t \notin \Omega_k$,

$$\mu_k(f, t) = \mu_k(f, t-1) + 2\pi S \nu_k(f, t), \tag{3}$$

where S is the hop size (in samples) of the STFT. In most papers using this phase model, such as [15, 16, 19], the authors assume that the mixtures are harmonic and stationary. In order to extend the validity of this technique to non-harmonic and non-stationary signals (which frequently occur in audio), we proposed in [13] to perform a local estimation (at each time frame) of instantaneous frequencies, by means of a Quadratic Interpolated FFT (QIFFT) performed on the spectra V_k. The frequency range is then decomposed into *regions of influence* [14] to ensure that the phase in a given channel is unwrapped with the appropriate frequency.

2.3. Main drawback

Assuming the magnitudes V_k, concentration parameters κ_k and prior phases μ_k are known, the source separation task consists in computing an estimator of the latent variables Z_k. In a probabilistic framework, the most natural estimators are given by the maximum likelihood, maximum a posteriori and posterior expectation estimators.

However, the computation of such estimators requires the knowledge of several PDFs, such as the likelihood, prior and posterior distributions. Despite some calculus efforts, we were not able to write the PDF of the mixture in closed-form. Thus, for obtaining these estimators, it is necessary to approximate the corresponding quantities with numerical schemes (for instance using Markov Chain Monte Carlo (MCMC) methods [24]). However, these techniques require several pieces of information on the variables' PDF (such as the likelihood), which are not available. Besides, they are computationally costly.

We thus propose to approximate the model (2) by a Gaussian model in which the moments of the sources are the same ones as in the original model (2). This approach enables us to keep the phase dependencies in a model which is fully tractable.

3. ANISOTROPIC GAUSSIAN MODEL

3.1. Mixture model

We approximate the Von Mises model (2) by a complex Gaussian model[1]:

$$X = \sum_k X_k \text{ with } X_k \sim \mathcal{N}(m_k, \gamma_k, c_k), \tag{4}$$

where $m_k = \mathbb{E}(X_k) \in \mathbb{C}$ is the mean of X_k, $\gamma_k = \mathbb{E}(|X_k - m_k|^2) \in \mathbb{R}_+$ is its variance and $c_k = \mathbb{E}((X_k - m_k)^2) \in \mathbb{C}$ is a *relation* term. The covariance matrix is:

$$\Gamma_k = \begin{pmatrix} \gamma_k & c_k \\ \bar{c}_k & \gamma_k \end{pmatrix}, \tag{5}$$

where $\bar{\cdot}$ denotes the complex conjugate. The PDF of a complex normal distribution $\mathcal{N}(m_k, \gamma_k, c_k)$ is:

$$p(X_k|m_k, \gamma_k, c_k) = \frac{1}{\pi\sqrt{\det(\Gamma_k)}} e^{-\frac{1}{2}(\underline{X}_k - \underline{m}_k)^H \Gamma_k^{-1} (\underline{X}_k - \underline{m}_k)}, \tag{6}$$

[1] Quite interestingly, such an approximation has been used in [25] where the mixture model was a sum of random variables with phase priors. This indicates that our approach is quite consistent with the technical issues that frequently arise in directional statistics modeling.

where $\underline{v} = \begin{pmatrix} v & \bar{v} \end{pmatrix}^T$ and $.^T$ (resp. $.^H$) denotes the transpose (resp. the conjugate tranpose). The keystone of our approach is that, in order to keep the phase dependencies, the moments are chosen such that they are the same ones in both models (2) and (4):

$$m_k = \mathbb{E}(X_k) = \mathbb{E}(Z_k), \tag{7}$$

$$\gamma_k = \mathbb{E}(|X_k - m_k|^2) = \mathbb{E}(|Z_k - m_k|^2), \tag{8}$$

$$c_k = \mathbb{E}((X_k - m_k)^2) = \mathbb{E}((Z_k - m_k)^2). \tag{9}$$

For a Von Mises random variable $\phi_k \sim \mathcal{VM}(\mu_k, \kappa_k)$, the n-th circular moment is, $\forall n \in \mathbb{Z}$:

$$\mathbb{E}(e^{in\phi_k}) = \frac{I_{|n|}(\kappa_k)}{I_0(\kappa_k)} e^{in\mu_k}. \tag{10}$$

Let us note

$$\lambda_k = \frac{I_1(\kappa_k)}{I_0(\kappa_k)}, \rho_k = \frac{I_2(\kappa_k)}{I_0(\kappa_k)} - \lambda_k^2, \tag{11}$$

and $\tilde{X}_k = V_k e^{i\mu_k}$ the estimated k-th component using the phase prior μ_k. Some simple algebra leads to:

$$m_k = \lambda_k \tilde{X}_k, \gamma_k = \left(1 - \lambda_k^2\right) V_k^2 \text{ and } c_k = \rho_k \tilde{X}_k^2. \tag{12}$$

The additive property of the Gaussian distribution family then implies that $X \sim \mathcal{N}(m_X, \gamma_X, c_X)$ with the following moments:

$$m_X = \sum_k m_k, \gamma_X = \sum_k \gamma_k, c_X = \sum_k c_k, \Gamma_X = \sum_k \Gamma_k. \tag{13}$$

3.2. MMSE estimator of the sources

The MMSE estimator of the sources is given by the posterior expectation of the components $\mathbb{E}(X_k|X)$. For Gaussian mixtures, this expectation is given by (see for instance [26]):

$$\hat{\underline{X}}_k = \underline{m}_k + \Gamma_k \Gamma_X^{-1} (\underline{X} - \underline{m}_X). \tag{14}$$

The set of estimators defined by (14) is conservative. Indeed, since $\Gamma_X = \sum_k \Gamma_k$, then $\sum_k \hat{\underline{X}}_k = \sum_k \underline{m}_k + (\underline{X} - \underline{m}_X) = \underline{X}$. Thus, this model preserves the overall energy of the mixture, which is not the case for the estimators $\tilde{X}_k$.

Such an estimator performs an interpolation between the prior estimate $\tilde{X}_k$ and the Wiener filtering estimate $G_k X$, where $G_k = \frac{V_k^2}{\sum_l V_l^2}$ is the traditional Wiener gain. Indeed, if $\forall k$, $\kappa_k \to 0$, then $\lambda_k \to 0$, then $\hat{X}_k \to G_k X$, which corresponds to the traditional Wiener filtering. This is coherent with the fact that for a null concentration parameter, the Von Mises distribution becomes equivalent to the uniform distribution. Then, the Gaussian model becomes isotropic, and in consequence, the MMSE estimator of the sources is given by the well-known Wiener filtering [3].

The proposed estimator (14) is thus expected to optimally exploit both the prior phase information and the mixture phase. Remarkably, an optimal combination of Wiener filtering and phase unwrapping estimates was proposed in [27], though it was restricted to mixtures of 2 sources in a speech enhancement framework.

3.3. Source separation procedure

The phase prior $\mu_k(f,t)$ can be computed by the phase unwrapping approach from the phase prior in the previous frame $\mu_k(f, t-1)$, as written in (3). However, a better approach seems to unwrap it from the MMSE-estimated phase $\angle \hat{X}_k(f, t-1)$, in order to avoid propagating the prior error.

We thus propose a procedure that is sequential over time frames: it consists in computing the phase prior and then the MMSE estimator in a given time frame before proceeding to the next frame. It is summarized in Algorithm 1.

Algorithm 1 Phase unwrapping informed source separation.

Inputs: Mixture $X \in \mathbb{C}^{F \times T}$,
$\forall k \in [\![1; K]\!]$: concentration parameters $\kappa_k \in \mathbb{R}_+^{F \times T}$, spectrograms $V_k \in \mathbb{R}_+^{F \times T}$, onset frames Ω_k and phases $\phi_k^o(f,t)$.
for $t = 1$ to $T-1$ **do**
  for $k = 1$ to K **do**
    if $t \in \Omega_k$ **then**
      Onset phases : $\forall f$, $\mu_k(f,t) = \phi_k^o(f,t)$.
    else
      $\mu_k(f,t)$ is unwrapped from $\angle \hat{X}_k(f, t-1)$.
    end if
  end for
  For each source k and channel f:
  Compute the prior estimate $\tilde{X}_k(f,t) = V_k(f,t) e^{i\mu_k(f,t)}$.
  Compute $\lambda_k(f,t)$ and $\rho_k(f,t)$ from (11).
  Compute $m_k(f,t)$, $\gamma_k(f,t)$ and $c_k(f,t)$ from (12).
  Compute $m_X(f,t)$, $\gamma_X(f,t)$ and $c_X(f,t)$ from (13).
  Compute $\Gamma_k(f,t)$ from (5) and $\Gamma_X(f,t)$ from (13).
  Compute the estimator $\hat{X}_k(f,t)$ from (14).
end for
Outputs: $\forall k \in [\![1; K]\!]$, $\hat{X}_k \in \mathbb{C}^{F \times T}$.

4. EXPERIMENTAL EVALUATION

In this section, we propose to experimentally assess the potential of the procedure described in Algorithm 1. We consider 100 music songs from the Demixing Secrets Database (DSD100), a remastered version of the database used for the SiSEC 2015 campaign [28]. The database is split into two sets of 50 songs: a learning database and a test database. Each song is made up of $K = 4$ sources: `bass`, `drums`, `vocals` and `other` (which may contain various instruments such as guitar, piano...).

The signals are sampled at $F_s = 44100$ Hz and the STFT is computed with a 92 ms long (4096 samples) Hann window, 75 % overlap and no zero-padding. Two scenarios are considered: an Oracle scenario, in which the magnitude spectrograms V_k are assumed to be known (i.e. equal to the ground truth), and a more realistic scenario, in which the spectrograms are estimated from the Oracle values by means of an NMF with Kullback-Leibler divergence [2], which uses 50 iterations of multiplicative update rules and a rank of factorization of 10. Note that this is not a fully blind scenario, since the NMFs are performed on the isolated spectrograms, but this will inform us about the performance of the methods when the spectrograms are no longer equal to the ground truth. The sets of onset frames Ω_k are detected with the MATLAB Tempogram Toolbox [29]. The mixture phase is given to the sources within onset frames as an input of Algorithm 1: $\forall k, f, \forall t \in \Omega_k$, $\phi_k^o(f,t) = \angle X(f,t)$. We consider a constant concentration parameter: $\forall (k,f,t)$, $\kappa_k(f,t) = \kappa$. The source separation quality is measured with the Signal to Distortion, Interference and Artifact Ratios (SDR, SIR and SAR) computed with the BSS EVAL toolbox [30]. Sound excerpts can be found on the companion website for this paper [31] to illustrate the experiments.

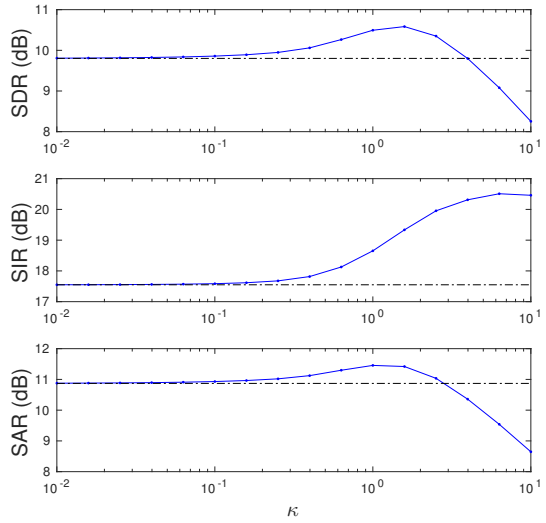

Fig. 2. Influence of the concentration parameter κ on the source separation quality in Algorithm 1 (solid lines) and comparison with Wiener filtering (dashed lines) in the Oracle scenario.

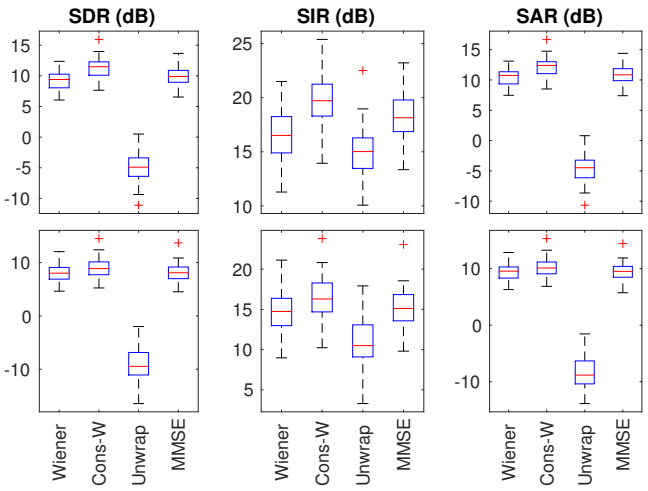

Fig. 3. Source separation performance for various methods on the DSD100 test dataset. Oracle (top) and estimated (bottom) magnitudes spectrograms.

4.1. Influence of the concentration parameter

We first study the impact of the concentration parameter κ on the separation quality. We test our procedure for various values of κ. We also compute the Wiener filtering estimates, as a comparison reference. Results averaged over the 50 songs composing the learning database are presented in Fig. 2 in the Oracle scenario.

We first observe that, for a certain range of values of κ, our approach leads to better results than the Wiener filtering technique. It shows that incorporating prior information about the phase in a source separation framework may increase the performance of the separation over a phase-unaware approach. When $\kappa \to 0$, we note that our approach becomes equivalent to Wiener filtering. Finally, the presence of SDR, SIR and SAR peaks suggests the existence of an optimal concentration parameter κ^* for this dataset, which corresponds to a compromise between excessively promoting the phase prior and only accounting for the mixture phase. Thus, this is consistent with our interpretation in Section 3.2.

The value $\kappa^* = 1.6$ seems to be a good compromise between these different indicators. Similar results are obtained in the non-Oracle scenario (leading to $\kappa^* = 1$), although the improvement over Wiener filtering is less important.

4.2. Source separation

We now consider the 50 songs that form the test database. We perform a source separation task with our procedure in both Oracle and non-Oracle scenarios. As comparison references, we also compute the Wiener filtering [3] and consistent Wiener filtering [11] estimates, and the components estimated using the phase prior only ($\tilde{X}_k = V_k e^{i\mu_k}$), denoted respectively by **Wiener**, **Cons-W** and **Unwrap**. The consistent Wiener filtering technique depends on a weight parameter that promotes the consistency constraint, which is learned beforehand on the learning database. The results are represented with box-plots in Fig. 3. Each box-plot is made up of a central line indicating the median of the data, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles, whiskers indicating the minimum and maximum values, and crosses representing the outliers.

We first observe that using the phase unwrapping prior only leads to poor results. Indeed, this technique neglects the phase of the mixture, then the prior error is propagated over time frames, leading to audible artifacts. In both Oracle and non-Oracle scenarios, the proposed estimator (denoted by **MMSE** in Fig. 3) leads to better results than **Wiener**, but slightly worse than **Cons-W** in terms of SDR, SIR and SAR. However, we perceptually observe that **Cons-W** tends to produce more artifacts in the `bass` and `drums` tracks than the proposed **MMSE** technique. Finally, it is important to note that **Cons-W** is computationally costly: for a 10 seconds excerpt, the separation is performed in 27 seconds with **Cons-W** vs 4 seconds with our estimator. The proposed approach then appears appealing for an efficient audio source separation task.

5. CONCLUSION

The model introduced in this paper is a promising tool for separating overlapping components in complex mixtures of audio signals in the TF domain. Incorporating prior information about the phase into a mixture model leads to a performance that is similar to Consistent Wiener filtering in terms of source separation quality, while significantly reducing the computational cost. In this paper, we used a prior obtained with the phase unwrapping algorithm, though the framework is very general and any phase prior could be used when computing the MMSE estimator of the sources.

The anisotropic Gaussian model then appears as an interesting approach to incorporate phase information in probabilistic mixture models, since it is fully tractable. While magnitudes values were assumed to be preliminary estimated in this study, future work will consist in introducing uncertainty about such a prior. Alternatively, one could model the magnitudes of the sources by Rayleigh random variables, whose dispersion parameters could be structured by means of an NMF model [3]. Such a model would be suitable for jointly estimating the magnitudes and phases of the components in a realistic audio source separation framework.

6. REFERENCES

[1] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] T. Virtanen, "Monaural Sound Source Separation by Non-negative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[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] T. Gerkmann, M. Krawczyk, and J. Le Roux, "Phase Processing for Single-Channel Speech Enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, March 2015.

[5] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, July 2016, Phase-Aware Signal Processing in Speech Communication.

[6] D. Griffin and J.S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, April 1984.

[7] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015.

[8] D. Gunawan and D. Sen, "Iterative Phase Estimation for the Synthesis of Separated Sources From Single-Channel Mixtures," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 421–424, May 2010.

[9] N. Sturmel and L. Daudet, "Iterative phase reconstruction of Wiener filtered signals," in *Proc. IEEE ICASSP*, Kyoto, Japan, March 2012.

[10] N. Sturmel and L. Daudet, "Informed Source Separation Using Iterative Reconstruction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 178–185, January 2013.

[11] J. Le Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 217–220, March 2013.

[12] R.J. McAuley and T.F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, August 1986.

[13] P. Magron, R. Badeau, and B. David, "Phase reconstruction of spectrograms with linear unwrapping: application to audio signal restoration," in *Proc. EUSIPCO*, Nice, France, August 2015.

[14] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[15] M. Krawczyk and T. Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, December 2014.

[16] P. Mowlaee, R. Saiedi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. of the International Conference on Spoken Language Processing*, Portland, OR, USA, September 2012.

[17] J. Bronson and P. Depalle, "Phase constrained complex NMF: Separating overlapping partials in mixtures of harmonic musical sources," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014.

[18] P. Magron, R. Badeau, and B. David, "Complex NMF under phase constraints based on signal modeling: application to audio source separation," in *Proc. IEEE ICASSP*, Shangai, China, March 2016.

[19] P. Mowlaee and J. Kulmer, "Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1521–1532, September 2015.

[20] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014.

[21] 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.

[22] Y. Agiomyrgiannakis and Y. Stylianou, "Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 775–786, May 2009.

[23] G.N. Watson, *A treatise on the theory of Bessel functions*, Cambridge university press, 1995.

[24] C. Andrieu, N. De Freitas, A. Doucet, and M.I. Jordan, "An introduction to MCMC for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[25] P. Beckmann, "Statistical distribution of the amplitude and phase of a multiply scattered field," *Journal of Research of the National Bureau of Standards*, vol. 66D, no. 3, pp. 231–240, May-June 1962.

[26] C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[27] M. Krawczyk and T. Gerkmann, "MMSE-optimal combination of wiener filtering and harmonic model based speech enhancement in a general framework," in *Proc. IEEE WASPAA*, New Paltz, NY, USA, 2015.

[28] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, pp. 387–395. Springer, 2015.

[29] P. Grosche and M. Müller, "Tempogram Toolbox: MATLAB tempo and pulse analysis of music recordings," in *Proc. ISMIR Conference*, Miami, FL, USA, October 2011.

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

[31] "Demo Webpage," "`http://perso.telecom-paristech.fr/magron/demos/demo_ICASSP2017.php`".