On modeling the STFT phase of audio signals with the von Mises distribution

Paul Magron, Tuomas Virtanen

ON MODELING THE STFT PHASE OF AUDIO SIGNALS WITH THE VON MISES DISTRIBUTION

Paul Magron, Tuomas Virtanen

Laboratory of Signal Processing, Tampere University of Technology, Finland
{firstname.lastname}@tut.fi

ABSTRACT

In this paper, we study statistical models for the phase of the short-time Fourier transform (STFT) of audio signals. STFT phase globally appears as uniformly distributed, which has led researchers in this field to model it as a uniform random variable. However, some information about the phase can be obtained from a sinusoidal model, which reveals its local structure. Therefore, we propose to model the phase with a von Mises (VM) random variable, which enables us to favor the sinusoidal model-based phase value. We estimate the distribution parameters and we validate this model on real audio data. In particular, we observe that both models (uniform and VM) are relevant from a statistical perspective but they convey different information about the phase (global vs. local). We also apply this VM model to an audio source separation task, where it outperforms previous approaches.

***Index Terms*—** STFT phase, von Mises distribution, mixtures of sinusoids, audio signal modeling, source separation

1. INTRODUCTION

Many audio signal processing techniques act on a time-frequency (TF) representation of the data such as the short-time Fourier transform (STFT), since the structure of such signals is more prominent in that domain. Much research has focused on the processing of STFT magnitude or power spectrograms in various applications such as automatic music transcription [1] and source separation [2].

However, processing STFT spectrograms results in discarding or not accounting for the phase information. For instance, in audio source separation applications, it is common to model the sources as circularly-symmetric random variables (e.g., Gaussian [3] or stable [4]), that is, not favoring any phase value. Alternatively, one can model and process spectrogram-like quantities only [2]. As a final processing stage, the phase of the isolated sources are retrieved by applying a Wiener-like filter, which assigns the phase of the original mixture to each extracted source. Those approaches are quite satisfactory in practice [2, 3], but it has been pointed out [5] that when sources overlap in the TF domain, it is responsible for residual interference and artifacts in the separated signals. This highlights the need for more refined phase models.

Even if the phase may appear as globally uniform [6], it holds some underlying structure that can be exploited. For instance, the sinusoidal model leads to explicit phase constraints in the TF domain [7]. These constraints have notably been exploited in speech enhancement [8] and source separation [5]. However, in a probabilistic framework, a uniform phase model does not allow us to exploit such a structure. To tackle this issue, some recent works proposed to model the phase with the von Mises (VM) distribution [9]. The VM distribution enables one to promote the sinusoidal model phase constraint, which has shown promising results in speech enhancement [10, 11] and source separation [12, 13].

In this paper, we propose to analyze the distribution of the STFT phase of audio signals from a statistical perspective. In particular, we establish that these two models (uniform and VM) are not contradictory, but rather complementary. Indeed, the uniform model originates from the underlying assumption that the phases across TF bins are independent and identically distributed (iid). From this perspective, the uniform model is statistically relevant, but it only conveys a *global* information. We show that a VM model that accounts for the underlying sinusoidal structure of audio is also statistically relevant, but it has the advantage of exploiting a *local* information about the phase. We introduce a simple procedure to estimate the VM model and we validate it experimentally on real audio data. We assess the potential of this technique for a source separation task, and we show that a proper choice of the VM distribution's parameters improves the separation quality over prior approaches.

The rest of this paper is structured as follows. In Section 2, we investigate on the uniformity assumption of the phase. In Section 3 we introduce and estimate the VM model that exploits the sinusoidal phase. Section 4 experimentally validates this model on real audio data and Section 5 shows its usefulness for a source separation task. Finally, Section 6 draws some concluding remarks.

2. IS THE PHASE REALLY UNIFORM?

Let $\mathbf{X} \in \mathbb{C}^{F \times T}$ be the STFT of a single-channel audio signal, where F and T are the number of frequency channels and time frames respectively. Its coefficient in the bin indexed by f and t is $x_{f,t}$, and the phase is $\phi_{f,t} = \angle x_{f,t}$, where $\angle$ denotes the complex argument.

2.1. A simple example

Let us first empirically observe the distribution of the phase of audio signals. As a toy example, we consider a piano piece from the MAPS database [14]. The signal is sampled at 8 kHz, and we compute its STFT with a 125 ms-long Hann window and 75 % overlap. We represent its spectrogram and phase in Fig. 1, as well as the histogram of the phase. Since the phase value in TF bins where there is no energy is not relevant, we compute the histogram of the phase only from TF points that belongs to Ω: this is the set of bins corresponding to local maxima of the magnitude spectrum and their neighboring bins, that is, the two frequency channels that surround each peak.

From this histogram, we observe that the phase appears as uniformly-distributed. Such an experiment can be reproduced with many real data corresponding to various instruments, with the same

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

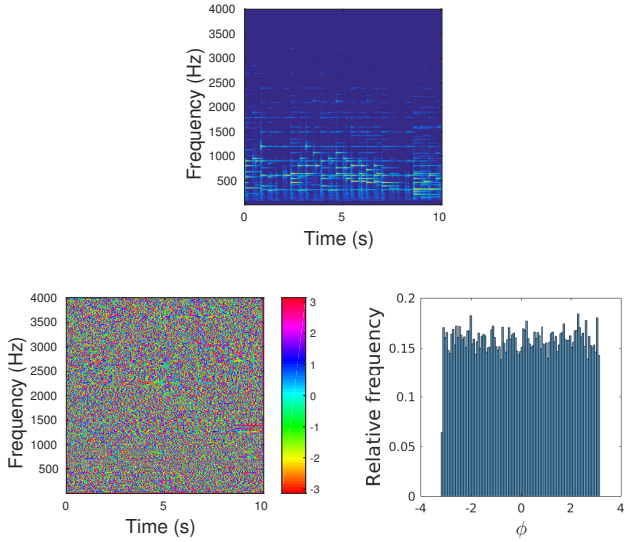

Fig. 1. Piano piece: spectrogram (top), phase (bottom left) and histogram of the phase (bottom right).

conclusion. This motivates the use of a uniform model for the phase, as frequently employed in the literature [6, 15].

2.2. Explicit phase constraint

The model of sum of sinusoids [16] is widely used to represent audio data [5, 8], notably piano signals such as the one used previously as toy example. It notably permits us to obtain explicit phase constraints in the STFT domain. Indeed, it can be shown [5] that for such signals, the phase follows the *phase unwrapping* relationship:

$$\phi_{f,t} \approx \phi_{f,t-1} + 2\pi l \nu_{f,t}, \tag{1}$$

where l is the hop size of the STFT and $\nu_{f,t}$ is the normalized frequency in channel f and time frame t. This relationship shows that if the phase in the previous time frame $\phi_{f,t-1}$ and the frequency $\nu_{f,t}$ are known, then the phase $\phi_{f,t}$ in the current time frame is completely determined (up to some error due to the frequency variation). This means that the phase holds some local structure, which is not accounted for in the uniform model, and is not revealed when globally observing the phase as in the previous experiment.

2.3. Statistical interpretation

Empirically observing a uniformly-distributed phase may appear contradictory with the fact that it holds some local structure. However, this uniform model is actually relevant from a statistical perspective. Indeed, in the toy example experiment conducted in Section 2.1, we *implicitly* assumed that the phases $\phi_{f,t}$ were iid. Plotting such a histogram reveals the underlying distribution of iid samples of a random variable, therefore this experiment was already assuming the iid property of the data, even though we did not formulate it explicitly.

Consequently, the uniform model is valid from a global perspective, provided an iid assumption of the phase. In order to account for a local structure of the phase such as (1), it is necessary to model the phases as non-uniform random variables, which also eliminates the need for assuming that the phases are iid.

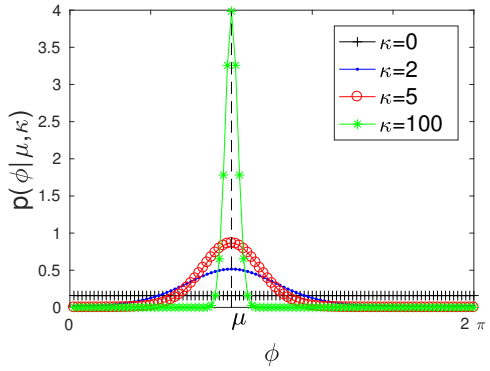

Fig. 2. Probability density function of the von Mises distribution.

3. VON MISES PHASE MODEL

3.1. Von Mises distribution

Recent phase models using the VM distribution [9] have been proposed, notably for speech enhancement [10, 11] and source separation [12, 13]. This suggest that the VM distribution is an appropriate tool for modeling the STFT phase of audio signals. The VM distribution, denoted $\mathcal{VM}(\mu, \kappa)$, depends on a location parameter $\mu \in [0; 2\pi[$ and a concentration parameter $\kappa \in [0; +\infty[$. Its probability density function, illustrated in Fig. 2, is given by:

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

where I_q is the modified Bessel function of the first kind of order q. The concentration parameter is analogous to the inverse of a variance parameter, which means that it quantifies how concentrated about the location parameter μ the distribution is. In particular, $\mathcal{VM}(\mu, 0)$ is the uniform distribution. Contrarily, if $\kappa \to +\infty$, it becomes equivalent to a Dirac delta function centered at μ.

In order to exploit the local structure of the phase, we propose to model the STFT phase in each TF bin with a VM random variable $\phi_{f,t} \sim \mathcal{VM}(\mu_{f,t}, \kappa)$, where the phase location parameter μ is given by the sinusoidal model (1):

$$\mu_{f,t} = \mu_{f,t-1} + 2\pi l \nu_{f,t}. \tag{3}$$

In particular, we remark that in this model, the phases are no longer identically distributed and they cannot be assumed independent.

3.2. Parameter estimation

To assess the validity of this model, we need to estimate its parameters. First, the frequencies $\nu_{f,t}$ are estimated by means of a quadratic interpolated FFT [17] performed on the peaks of the log-spectra of the sources at each time frame, in order to account for slow variations of the frequencies [5]. Given those frequencies and the phase in the previous time frame, we obtain estimates $\hat{\mu}_{f,t}$ of the location parameters according to (3). Then, we define the centered variables:

$$\psi_{f,t} = \phi_{f,t} - \hat{\mu}_{f,t}. \tag{4}$$

Those variables are independent given $\hat{\mu}_{f,t}$ and identically distributed: $\forall f, t, \psi_{f,t} \sim \mathcal{VM}(0, \kappa)$. We then propose to estimate the concentration parameter κ in a maximum likelihood sense [18]. The

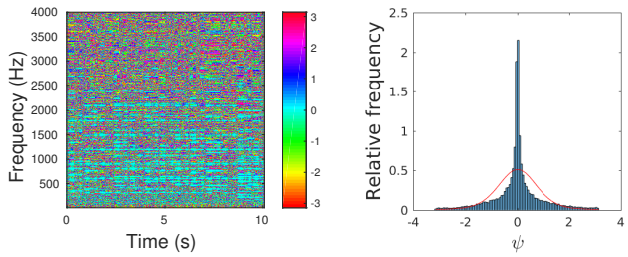

Fig. 3. Centered phases for the same example as in Fig. 1: phase (left) and histogram (right), where the solid line represents the fitted VM distribution.

log-likelihood is given by:

$$\begin{aligned}\mathcal{L}(\kappa) &= \sum_{(f,t)\in\Omega} \log p(\psi_{f,t}|\hat{\mu}_{f,t},\kappa) \\ &= \sum_{(f,t)\in\Omega} \kappa\cos(\psi_{f,t}) - \log(2\pi) - \log(I_0(\kappa)) \\ &= -|\Omega|(\log(2\pi) + \log(I_0(\kappa))) + \kappa \sum_{(f,t)\in\Omega} \cos(\psi_{f,t}).\end{aligned}$$

where $|\Omega|$ denotes the cardinality of the set Ω defined in Section 2.1. Setting its derivative with respect to κ to 0 leads to:

$$\frac{I_1(\kappa)}{I_0(\kappa)} = \frac{1}{|\Omega|} \sum_{(f,t)\in\Omega} \cos(\psi_{f,t}), \tag{5}$$

and solving the implicit equation (5) yields an estimate of κ. Since the function $\kappa \to \frac{I_1(\kappa)}{I_0(\kappa)}$ is monotonic and concave on $[0, +\infty[$, this equation can be efficiently solved with numerical methods (here, we simply used the `fzero` Matlab function).

4. EXPERIMENTAL VALIDATION

In this section, we assess the validity of the VM model for representing real audio data.

4.1. Validation of the model

We consider 30 piano pieces from the MAPS database [14] and 6 guitar pieces from the IDMT-SMT-GUITAR database [19]. For each dataset, we concatenate all the music excerpts into one long signal. The signals are sampled at 8 kHz and the STFT is computed with a 125 ms-long Hann window and 75 % overlap. We estimate the corresponding VM distribution parameters accordingly to the procedure presented in Section 3.

We illustrate the results on the piano piece used in Section 2.1. We plot the centered variables $\psi_{f,t}$ and their histogram in Fig. 3 (as in Fig. 1, this histogram only uses TF points of significant energy). We remark that the variables exhibit a VM trend, which confirms the appropriateness of this distribution for modeling the phase of audio signals in the STFT domain.

We obtain overall similar results for both databases. We note that the concentration parameter is greater for piano sounds ($\kappa \approx 2$), which are known to be quite accurately represented with mixtures of slowly-varying sinusoids [20], than for guitar signals ($\kappa \approx 1.7$). Since κ quantifies how close to the location parameter (given here by a sinusoidal model) the phases are distributed on average, κ quantifies the *sinusoidality* of the data.

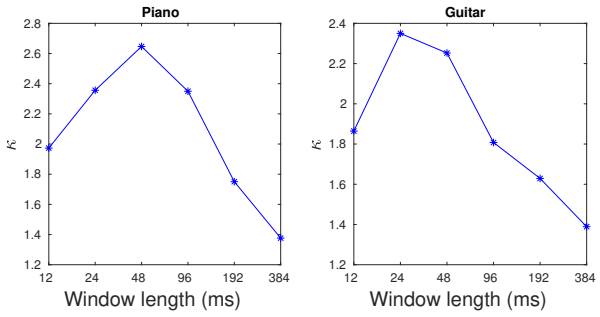

Fig. 4. Influence of the window length on the concentration parameter κ for piano and guitar signals.

4.2. Impact of the window length

The accuracy of the sinusoidal model does not only depend on the nature of the data, but also on the STFT parameters, such as the window length [5]. Indeed, the time and frequency resolutions of the STFT strongly impact the local stationarity and slow-variation assumptions used to derive (1). Therefore, we compute the concentration κ for several lengths of the analysis window, ranging from 12 ms to 384 ms.

The results provided in Fig. 4 suggest that there exist an optimal window length (which depends on the nature of the signals) for which the sinusoidality of the data is maximized: the phase follows the sinusoidal model more closely with this value than with another, which may be an interesting criterion for tuning the STFT parameters (*cf.* Section 5.1).

4.3. Discussion

The two models (uniform and VM) are both statistically relevant as shown in Fig. 1 and 3. The difference is that the uniform model is based on the iid assumption of the phases across time and frequency, while in the VM model the iid variables are the centered phases conditionally to an estimate of the sinusoidal location parameters.

Consequently, both models are adapted for representing the STFT phase of audio signals, but they do not carry the same type of information. The uniform model carries a *global* information on the overall distribution of the phase. The VM model accounts for some *local* property, and makes it possible to exploit a local phase model, such as the sinusoidal phase constraint (1).

Therefore, the model should be chosen accordingly to the scenario and target application. For instance, in audio source separation, refined modeling of the phase is important if the sources are strongly overlapping in the TF domain, but a uniform model can be sufficient otherwise [12].

5. APPLICATION TO AUDIO SOURCE SEPARATION

Finally, we propose to apply this VM framework to an audio source separation task. In [12], we introduced a VM-based probabilistic source model in order to account for the sinusoidal local structure of the phase. It resulted in estimating the sources through an *anisotropic* Wiener (AW) filtering technique, which optimally combines the mixture's phase and the sinusoidal model, and outperformed the traditional phase-unaware Wiener filter [3]. However, the concentration parameter was the same for all sources and it was selected by maximizing a set of objective criteria, that is, the

Table 1. Source separation performance averaged over the DSD100 test dataset. Higher is better.

	Bass			Drums			Other			Vocals			Average		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Wiener	7.0	15.7	7.7	8.9	19.2	9.5	7.0	17.1	7.6	11.3	24.4	11.6	8.5	19.1	9.1
AW-unif	8.3	17.7	8.9	9.6	**22.4**	9.9	8.1	19.2	8.6	12.2	**27.1**	12.3	9.5	21.6	9.9
AW-var	**8.5**	**19.2**	**9.0**	**9.9**	22.1	**10.2**	**9.3**	**19.4**	**8.8**	**12.2**	26.7	**12.4**	**9.7**	**21.9**	**10.1**

signal-to-distortion, signal-to-interference, and signal-to-artifact ratios (SDR, SIR, and SAR) [21]. Here, we rather propose to use a different κ for each source, and to estimate it with the framework introduced in this paper.

5.1. Setup

We consider 100 music song excerpts from the DSD100 database [22]. Each excerpt, sampled at 44100 Hz, is 10 seconds-long and is made up of 4 sources: `bass`, `drums`, `vocals` and `other`. The dataset is split into two sets of 50 songs: the learning and test sets.

To determine the STFT window length, we apply the procedure described in Section 4.2. The sinusoidality of the data is maximal for windows of 48 ms and 92 ms. Therefore, we compute the STFT with a 92 ms-long Hann window and 75 % overlap, since this window length leads to better overall separation results than a 48 ms window.

The sources' magnitudes are assumed known (i.e., equal to their values before mixing), so we only investigate on the impact of phase recovery on source separation. The performance of the AW filter in the presence of magnitude estimation errors has been studied in [12].

Source separation quality is measured with the SDR, SIR, and SAR [21] expressed in dB, where only a rescaling (not a refiltering) of the reference is allowed.

5.2. Learning the concentration parameters

Firstly, we estimate κ for each isolated source on the 50 songs that form the learning set and we plot the results in Fig. 5. We observe that there is a high variability of κ for the `bass` and `other` tracks. This can be explained by a high variability of the sounds themselves in these tracks: for instance, in the `bass` track, the instrument can be acoustic, electric, fretless or electronic, and the playing technique can be finger picking or slap. The lowest concentration parameters are obtained for the `drums` tracks, which is consistent with the non sinusoidal nature of those signals.

Interestingly, we remark that the average of κ over all sources and over the dataset is 1.6, which is very close to the value obtained with our previous learning approach [12]. The resulting separation procedures are then expected to yield quite similar results, however they differ regarding two aspects. Firstly, the approach presented in this paper allows us to choose an optimal concentration parameter for each individual instrument. Secondly, it is significantly less computationally demanding than the previous technique, which required performing the whole separation for many values of κ before picking the best value.

5.3. Comparison to other methods

As baseline methods, we test the phase-unaware Wiener filter [3] and AW with a uniform concentration parameter for all sources, which is learned as in [12], that is, such that is maximizes the average SDR, SIR and SAR. This method will be referred to as AW-unif. The method proposed here is the AW filter with a specific concentration parameter for each source, which value is the median obtained at the learning stage (*cf.* previous Section). This method will be referred to as AW-var.

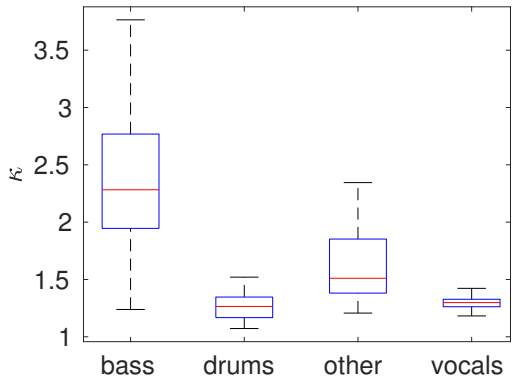

Fig. 5. Estimated concentration parameter κ for each source of the DSD100 learning dataset. Each box-plot is made up of a central line indicating the median, upper and lower box edges indicating the 1^{st} and 3^{rd} quartiles and whiskers indicating the extrema.

The results are reported in Table 1. The proposed AW-var approach improves all the metrics compared to AW-unif in the `bass` and `other` tracks. It also reduces the artifacts and distortion in the other tracks, but AW-unif yields more interference rejection in the `drums` and `vocals` tracks. The proposed approach results in a better overall separation, as attested by the increase in average SDR, SIR and SAR compared to AW-unif.

It should be noted that the tracks for which improvement is not observed for all the criteria (`drums` and `vocals`) are the tracks for which κ has the lowest variability (*cf.* Fig. 5). Therefore, a potential future research direction is to account for this intra-track variability by modeling κ as a random variable in a hierarchical model [23].

6. CONCLUSION

In this paper, we have shown that both a uniform and a VM model for the STFT phase of audio signals are statistically relevant, even though they do not convey the same type of information. In particular, the VM distribution is an interesting tool for accounting for local constraints of the phase, such as a property that arises from a sinusoidal model. This model has been validated on real audio data, from which we could interpret the concentration parameter of the VM distribution as a measure of the sinusoidality of the data. This model and the corresponding estimation technique has been shown useful for audio source separation, and may further be exploited in more sophisticated separation techniques [13, 24]. Alternatively, measuring the sinusoidality of audio signals could be useful for harmonic/percussive instrument recognition.

7. REFERENCES

[1] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2003.

[2] 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.

[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 and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[5] 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*, vol. 26, no. 6, pp. 1095–1105, June 2018.

[6] 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.

[7] M. Krawczyk and T. Gerkmann, "STFT phase improvement for single channel speech enhancement," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, September 2012.

[8] 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.

[9] 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.

[10] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4199–4208, August 2014.

[11] J. Kulmer and P. Mowlaee, "Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015.

[12] P. Magron, R. Badeau, and B. David, "Phase-dependent anisotropic Gaussian model for audio source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

[13] P. Magron and T. Virtanen, "Bayesian anisotropic Gaussian model for audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[14] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, August 2010.

[15] R. Mitchell Parry and Irfan Essa, "Phase-aware non-negative spectrogram factorization," in *Proc. of International Conference on Independent Component Analysis and Signal Separation*, September 2007.

[16] R. J. McAuley and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, August 1986.

[17] 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.

[18] S. Calderara, A. Prati, and R. Cucchiara, "Mixtures of von Mises distributions for people trajectory shape analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 457–471, April 2011.

[19] C. Kehling, Abeßer J., Dittmar C., and Schuller G., "Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, September 2014.

[20] W.M. Szeto and K.H. Wong, "Source separation and analysis of piano music signals using instrument-specific sinusoidal model," in *Proc. of International Conference on Digital Audio Effects (DAFx)*, September 2013.

[21] 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.

[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] G.M. Allenby, P.E. Rossi, and R.E. McCulloch, "Hierarchical Bayes models: A practitioners guide," January 2005.

[24] 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.