Open access • Proceedings Article • DOI:10.1109/ICASSP.2016.7471650

Common fate model for unison source separation — Source link

Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler ...+1 more authors

Institutions: French Institute for Research in Computer Science and Automation, Institut Mines-Télécom

Published on: 20 Mar 2016 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Musical instrument, Spectrogram, Source separation, Discrete Fourier transform and Amplitude modulation

Common Fate Model for Unison source Separation

Fabian-Robert Stöter, Antoine Liutkus, Roland Badeau, Bernd Edler, Paul Magron

COMMON FATE MODEL FOR UNISON SOURCE SEPARATION

Fabian-Robert Stöter[⋆] *Antoine Liutkus*[†] *Roland Badeau*[‡] *Bernd Edler*[⋆] *Paul Magron*[‡]

[⋆] International Audio Laboratories Erlangen[∗]
[†] Inria, speech processing team, Villers-les-Nancy, France
[‡] Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France

ABSTRACT

In this paper we present a novel source separation method aiming to overcome the difficulty of modelling non-stationary signals. The method can be applied to mixtures of musical instruments with frequency and/or amplitude modulation, e.g. typically caused by vibrato. It is based on a signal representation that divides the complex spectrogram into a grid of patches of arbitrary size. These complex patches are then processed by a two-dimensional discrete Fourier transform, forming a tensor representation which reveals spectral and temporal modulation textures. Our representation can be seen as an alternative to modulation transforms computed on magnitude spectrograms. An adapted factorization model allows to decompose different time-varying harmonic sources based on their particular common modulation profile: hence the name *Common Fate Model*. The method is evaluated on musical instrument mixtures playing the same fundamental frequency (unison), showing improvement over other state-of-the-art methods.

Index Terms— Sound source separation, Common Fate Model, Non-Negative tensor factorization.

1. INTRODUCTION

Sound source separation continues to be a very active field of research [1] with a variety of applications. Many recent contributions are based on the popular non-negative matrix factorization (NMF). The way NMF factorizes a spectrogram matrix into frequency and activation templates makes it possible to easily design algorithms in an intuitive way. At the same time, it provides a rank reduction, needed to decompose mixtures into their source components. In the past, many NMF-based source separation methods have been developed [2–4]. Expanding the NMF to tensors allows to incorporate more complex models, useful in many applications like multichannel separation. Extensions to NMF such as shift-invariance or convolutions were carried over to non-negative tensor (NTF) based algorithms [5–9]. These approaches, relying on decomposing mixtures of musical instruments, work well when certain assumptions hold to be true. One is that spectral harmonics only partially overlap. However, when two sources share the same fundamental frequency, almost all partials do overlap, making it difficult for NMF-based algorithms to learn unique templates. Another assumption is that all spectral and temporal templates semantically correspond to musical notes, forming a dictionary of musically meaningful atoms. This does not hold for instruments with time-varying fluctuations. These effects can typically be found in musical instruments like strings and brass, when played with vibrato. In a setting where two musical instruments with vibrato play in unison, both assumptions could break, which makes it a challenging scenario [10]. When processing such mixtures with a representation based on a standard NMF and the magnitude spectrogram, it is hard to model the sources with only a few spectral templates. Instead of increasing the number of templates per source, Hennequin proposes [11] frequency-dependent activation matrices by using a source/filter-based model. Since the vibrato does not only cause frequency modulation (FM) but also amplitude modulation (AM), so-called modulation spectra can be used to identify the modulation pattern. This is often calculated by taking the Fourier transform of a magnitude spectrum. Thus, the *modulation spectrogram* has already gathered much attention in speech recognition [12, 13] and classification [14, 15]. Barker and Virtanen [16] were the first to propose a modulation tensor representation for single channel source separation. This allows to elegantly apply factorization on the tensor by using the well known PARAFAC/CANDECOMP (CP) decomposition.

In this work we introduce a novel tensor signal representation which additionally exploits similarities in the frequency direction. We can therefore make use of dependencies between modulations of neighbouring bins. This is similar to the recently proposed High-Resolution Nonnegative Matrix Factorization model that accounts for dependencies in the time-frequency plane (HR-NMF [17]). In short, HR-NMF models each complex entry of a time-frequency transform of an audio signal as a linear combination of its neighbours, enabling the modelling of damped sinusoids, along with an independent innovation. This model was generalized to multichannel mixtures in [18, 19] and was shown to provide considerably better oracle performance for source separation than alternative models in [20]. Indeed, even though some variational approximations were introduced in [21] to strongly reduce their complexity, those algorithms are often demanding for practical applications. In this paper, we propose to relax some assumptions of HR-NMF in the interest of simplifying the estimation procedure. The core idea is to divide the complex spectrogram into modulation patches in order to group common modulation in time and frequency direction. We call this the *Common Fate Model* (CFM), borrowing from the Gestalt theory, which describes how human perception merges objects that move together over time. Bregman [22] described the Common Fate theory for auditory scene analysis as the ability to group sound objects based on their common motion over time, as occurs with frequency modulations of harmonic partials. As outlined by Bregman, the human ability to detect and group sound sources by small differences in FM and AM is outstanding. Also, it turns out that humans are especially sensitive to modulation frequencies around 5 Hz, which is the typical vibrato frequency that many musicians produce naturally.

∗ A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

This work was partly supported by the research programmes EDiSon3D (ANR-13-CORD-0008-01) and KAMoulox (ANR-15-CE38-0003-01) funded by ANR, the French State agency for research.

2. COMMON FATE MODELLING

2.1. The Common Fate Transform

Let $\tilde{x}$ denote a single channel audio signal. Its Short-Term Fourier Transform (STFT) is computed by splitting it into overlapping frames, and then taking the discrete Fourier transform (DFT) of each one[1]. The resulting information is gathered into an $N_\omega \times N_\tau$ matrix written X, where N_ω is the number of frequency bands and N_τ the total number of frames. In this study, we will consider the properties of another object, built from X, which we call the Common Fate Transform (CFT). It is constructed as illustrated in Figure 1. We split the STFT X into overlapping rectangular $N_a \times N_b$ patches, regularly spaced over both time and frequency. Then, the 2D-DFT of each patch is computed[2]. This yields an $N_a \times N_b \times N_f \times N_t$ tensor we write x, where N_f and N_t are the vertical and horizontal positions for the patches, respectively.

As can be seen, the CFT is basically a further short-term 2D-DFT taken over the standard STFT X. One of the main differences compared to modulation spectrograms is that the CFT is computed using the complex STFT X, and not a magnitude representation such as $|X|$. As we will show, this simple difference has many interesting consequences, notably that the CFT is invertible: the original waveform $\tilde{x}$ can be exactly recovered by cascading two classical overlap-add procedures. Another difference is that the patches span several frequency bins, *i.e.* we may have $N_a > 1$. This contrasts with the conventional modulation spectrogram, that is usually defined using one frequency band only.

2.2. A Probabilistic Model for the CFT

When processing an audio signal $\tilde{x}$ for source separation, it is very common to assume that all time-frequency (TF) bins of its STFT are independent [23–26]. This is often the consequence of two different assumptions. The first one is to consider that all frames are independent, thus leading to the independence of all entries of the STFT that do not belong to the same column. The second one is related to the notion of stationarity: roughly speaking, the Fourier transform is known to decompose stationary signals into independent components, whether these signals be Gaussian (see, e.g. [26]) or, more generally, harmonisable α-stable [27]. As a consequence, when the signals are assumed to be *locally stationary*, it is theoretically sound to assume that all the entries of their STFT are independent.

Still, both assumptions can only be considered as approximations. First, adjacent frames are obviously not independent, notably because of the overlap between them. Second, the stationarity assumption is only approximate in practice, especially when impulsive elements are found in the audio, leading to strong dependencies among the different frequency bins. Let $\{X_{ft}\}_{f,t}$ denote all the $N_a \times N_b$ patches taken on the STFT to compute the CFT, as depicted in Figure 1. The probabilistic model we choose is the combination of *four* different assumptions made on the distribution of these patches.

1. All patches are independent. Just as the classical locally stationary model [26] assumes independence of overlapping frames, we assume here independence of overlapping patches. Due to the overlap between them, this assumption is an approximation, and one may wonder what the advantage is of dropping independent frames for independent patches. The answer lies in the fact that the latter permits us to model phase dependencies between neighbouring STFT entries, and also to model much longer-term dependencies, as required for instance by deterministic damped or frequency-modulated sinusoidal signals.

2. Each patch is *stationary*: its distribution is assumed invariant under translations in the TF plane. This is where we do not assume independence, but on the contrary expect dependencies among neighbouring STFT entries. Our approach assumes this happens in a way that only depends on the relative positions in the TF plane. It can easily be shown that mixtures of damped sinusoids have this property. Assuming stationarity not only over time but over both time and frequency also permits us to naturally account for mixtures of frequency-modulated sounds. In short, we assume that throughout each patch, we observe one coherent STFT "texture". The difference with the HR-NMF model is that we have independent and identically distributed (i.i.d.) innovations for one given patch, whereas HR-NMF model has more variability and permits heteroscedastic innovations. However, taking overlapping patches somehow compensates for this limitation.

3. The joint distribution of all entries of each patch is α-stable [28]. α-stable distributions are the only ones that are stable under additions, *i.e.* such that sums of α-stable random variables (r.v.) remain α-stable. They notably comprise the Gaussian and Cauchy distribu-

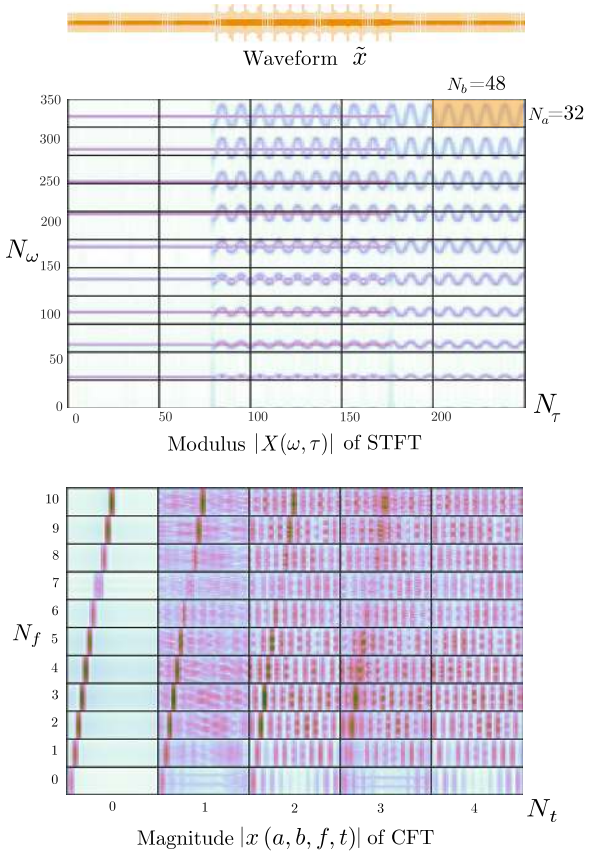

Fig. 1. Common Fate Transform. For convenience, the splitting of the STFT into patches has been depicted without overlap, but overlapping patches are used in practice.

[1] Since the waveform $\tilde{x}$ is real, the Fourier transform of each frame is Hermitian. In the following, we assume that the redundant information has been discarded to yield the STFT.

[2] Note that since each patch is complex, its 2D-DFT is not Hermitian, thus all its entries are kept.

tions as special cases when $\alpha = 2$ and $\alpha = 1$, respectively.

4. Each patch is harmonisable, *i.e.* is the inverse Fourier transform of a complex random measure with independent increments. In other words, all entries of the Fourier transform of each patch are assumed to be asymptotically independent, as the size of the patch gets larger. This rather technical condition, often tacitly made in signal processing studies, permits efficient processing in the frequency domain.

Under those four assumptions, all entries of the CFT x are independent (assumptions 1 and 2), and each one is distributed with respect to a complex isotropic α-stable distribution, noted $S\alpha S_c$ (assumptions 3 and 4[3]):

$$x\left(a,b,f,t\right) \sim S\alpha S_c\left(P^{\alpha}\left(a,b,f,t\right)\right), \tag{1}$$

where P^{α} is a nonnegative $N_a \times N_b \times N_f \times N_t$ tensor that we call the *modulation density*. When $\alpha = 2$, (1) corresponds to the classical isotropic complex Gaussian distribution and the entries of P^{α} are homogeneous to variances. In the general case, it can basically be understood as the energy found at (a, b) for patch (f, t), just like more classical (fractional) power spectral densities describe the spectro-temporal energy content of the STFT of a locally stationary signal.

2.3. Signal Separation

Now, let us assume that the observed waveform is actually the sum of J underlying sources $\{\tilde{s}_j\}_{j=1,\ldots,J}$. Due to the linearity of the CFT, this can be expressed in the CFT domain as:

$$\forall\left(a,b,f,t\right), x\left(a,b,f,t\right) = \sum\nolimits_j s_j\left(a,b,f,t\right).$$

If we adopt the α-stable model presented above for each source and use the stability property, we have:

$$x\left(a,b,f,t\right) \sim S\alpha S_c\left(\sum\nolimits_j P_j^{\alpha}\left(a,b,f,t\right)\right),$$

where P_j^{α} is the modulation density for source j. If these objects are known, it can be shown that each source can be estimated in a maximum a posteriori sense from the mixture as:

$$\mathbb{E}\left[s_j\left(a,b,f,t\right) \mid \{P_j^{\alpha}\}_j, x\right] = \frac{P_j^{\alpha}(a,b,f,t)}{\sum_{j'} P_{j'}^{\alpha}(a,b,f,t)}\, x\left(a,b,f,t\right) \tag{2}$$

which we call the fractional α-Wiener filter in [27]. The resulting waveforms are readily obtained by inverting the CFT. As can be seen, we now need to estimate the modulation densities $\{P_j^{\alpha}\}_j$ based on the observation of the mixture CFT x, similarly to the estimation of the sources' (fractional) Power Spectral Densities (α-PSD) in source separation studies.

2.4. Factorization Model and Parameter Estimation

In order to estimate the sources' modulation densities, we first impose a factorization model over them, so as to reduce the number of parameters to be estimated. In this study, we set:

$$P_j^{\alpha}\left(a,b,f,t\right) = A_j\left(a,b,f\right) H_j\left(t\right), \tag{3}$$

where A_j and H_j are $N_a \times N_b \times N_f$ and $N_t \times 1$ nonnegative tensors, respectively. We call this a *Common Fate Model*. Intuitively, A_j is a modulation density template that is different for each frequency band f, and that captures the long term modulation profile of source j around that frequency. Then, H_j is an activation vector that indicates the strength of source j on the patches located at temporal position t. Learning those parameters can be achieved using the conventional Nonnegative Matrix Factorization methodology (NMF, see e.g. [25, 29, 30] for an overview and [31] for the fitting of $S\alpha S_c$ parameters), except that it is applied to the CFT instead of the STFT, and that the particular factorization to be used is (3).

[3]This result is the direct generalization of [28, th. 6.5.1] to multi-dimensional stationary processes.

Algorithm 1 Fitting NMF parameters of the nonnegative CFM (3).

With $v^{\alpha} = |x|^{\alpha}$ and always using the latest parameters available for computing $\hat{P}^{\alpha}\left(a,b,f,t\right) = \sum_{j=1}^{J} A_j\left(a,b,f\right) H_j\left(t\right)$, iterate:

$$A_j\left(a,b,f\right) \leftarrow A_j\left(a,b,f\right) \frac{\sum_t v^{\alpha}(a,b,f,t)\hat{P}^{\alpha}(a,b,f,t)^{\cdot(\beta-2)} H_j(t)}{\sum_t \hat{P}^{\alpha}(a,b,f,t)^{\cdot(\beta-1)} H_j(t)}$$

$$H_j\left(t\right) \leftarrow H_j\left(t\right) \frac{\sum_{a,b,f} v^{\alpha}(a,b,f,t)\hat{P}^{\alpha}(a,b,f,t)^{\cdot(\beta-2)} A_j(a,b,f)}{\sum_{a,b,f} \hat{P}^{\alpha}(a,b,f,t)^{\cdot(\beta-1)} A_j(a,b,f)}.$$

Due to space constraints, we cannot detail the derivations of the fitting strategy. In essence, it amounts to estimating the parameters $\{A_j, H_j\}$ so that the modulus of the CFT, raised to the power α, is as close as possible to $\sum_j P_j^{\alpha}$, with some particular cost function as a data-fit criterion, called a β-divergence and which includes Euclidean, Kullback-Leibler and Itakura-Saito as special cases [32]. As usual in such nonnegative models, each parameter is updated in turn, while the others are kept fixed. We provide the multiplicative updates in Algorithm 1. After a few iterations, the parameters can be used in (2) to separate the sources.

3. EXPERIMENTS

In this section, we present separation experiments utilizing CFM and we compare it with other methods.

3.1. Synthetic Example

To illustrate the CFT representation we processed a mixture consisting of two sinusoidal sources. One source is a pure sine wave of fundamental frequency 440 Hz whereas the other is frequency modulated by a sinusoid of 6.3 Hz. In the first step an STFT with a DFT-length of 1024 samples and a hop-size of 256 samples was processed at a sample rate of 22.05 kHz. Patches of size $(N_a, N_b) = (32, 48)$ (not respecting overlaps) were then taken from the STFT output. Figure 1 in Section 2.1 then shows the Common Fate Transform for the mixture as described in Section 2. One can see that the CFT representation shows distinct patterns across time, suggesting that the factorization is able to separate the sources.

3.2. Objective Evaluation on Unison Instrument Mixtures

For an evaluation of the method, we selected five musical instruments' samples, all featuring vibrato: violin, cello, tenor sax, English horn, and flute. It is important to note that vibrato techniques differ between instruments: whereas the English horn and the flute only produce a very subtle modulation, the violin and tenor sax have powerful frequency modulations with a higher modulation frequency as well as a higher modulation index. The signals have each

Method	Description	Signal Representation	Factorization Model
CFM	Common Fate Model	STFT $\rightarrow$ Grid Slicing $\rightarrow$ 2D-DFT	$V(a,b,f,t) = P(a,b,f) \times H(t)$
NMF	[4] w/o add. constraints	STFT	$V(f,t) = W(f) \times H(t)$
HR-NMF	High Resolution NMF model [20]	Output of any filterbank (STFT, MDCT, ...)	Subband AR filtering of NMF excitation
MOD	[16] using DFT filterbank	STFT $\rightarrow |\dots| \rightarrow$ STFT along each bin	$V(f,m,t) = W(f) \times A(m) \times H(t)$
CFMM	Common Fate Magnitude Model	STFT $\rightarrow |\dots| \rightarrow$ Grid Slicing $\rightarrow$ 2D-DFT	$V(a,b,f,t) = P(a,b,f) \cdot H(t)$
CFMMOD	CFMM with $a = 1$	STFT $\rightarrow |\dots| \rightarrow$ Grid Slicing $\rightarrow$ 2D-DFT	$V(a,b,f,t) = P(a,b,f) \cdot H(t)$

Table 1. Overview of the evaluated algorithms

been generated by rendering C4 (261.63 Hz) notes in a state-of-the-art software sampler[4]. All samples last about three seconds. We then generated a combination of ten mixtures of two instruments each, each one generated with a simple SourceA — SourceB — (SourceA + SourceB) scheme. Data were encoded in 44.1 kHz / 16 bit. For evaluation, we compared separation performance of six different methods, summarized in Table 1:

CFM For the CFM model, we took an STFT with frames of 1024 samples and a hop-size of 512 samples. The resulting complex spectrogram was then split into a grid of patches of size $(N_a, N_b) = (4, 64)$, each having a half-window overlap in both dimensions. For all experiments α and β were set to 1.

MOD We implemented a modified version of [16] where for the sake of comparability, we used a STFT instead of a gammatone filterbank. A DFT length of 1024 and a hop-size of 512 samples were chosen. After taking the magnitude value, a second STFT of size 32 and hop-size 16 samples was computed for each frequency.

CFMMOD We selected patch sizes of $(N_a, N_b) = (1, 64)$ and modified the representation so that the magnitude spectrogram was used before computing the 2D-DFT. This permits to compare the advantage of our proposed factorization model (3) over MOD, when using the same kind of energy-modulation representation in both cases.

CFMM For comparing the influence of computing modulations over complex STFT or magnitude spectrograms, we tried our factorization model when the magnitude of the STFT is taken before 2D-DFT, with patches of the same size as for the CFM method.

NMF We took a standard NMF based method [4]. We highlight that taking a spectrogram with frames of length 1024 would not make a fair comparison, because the CFM model actually results in a larger frequency resolution. Therefore a comparable NMF is based on an STFT of DFT-length 32768.

HR-NMF See description in [20].

All factorizations ran for 100 iterations and were repeated five times. We chose $j = (2 \dots 6)$ components for each factorization. For $j > 2$ we used oracle clustering to show the upper limit of SDR which can be achieved.

We ran the performance evaluation by using BSSeval [33]. The results of Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifacts Ration (SAR) are depicted in Figure 2. Results indicate that the CFM model performs well in all measures. However, in terms of SIR the results of HR-NMF are better than CFM method. The results for CFMMOD indicate the positive influence of the CFM factorization compared to [16]. The results of CFMM indicate that the complex CFT lead to better results. NMF did perform surprisingly well, which may only hold for our test set, where each source is active for a long period. This results in a cyclic stationary vibrato, revealing spectral side lobes at such a high resolution. With more than one component per source, the results of CFM do improve, but it can be seen that more than two components ($j = 4$) will not increase the SDR values. The separation results and a full Python implementation of the CFM algorithm can be found on the companion website for this paper[5].

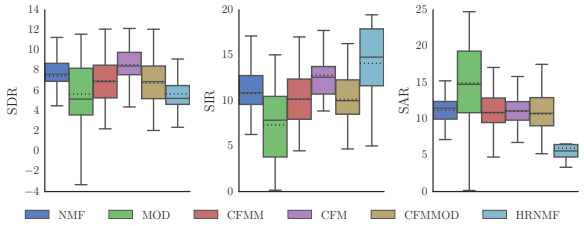

Fig. 2. Boxplots of BSS-Eval results of the unison dataset. Solid/dotted lines represent medians/means.

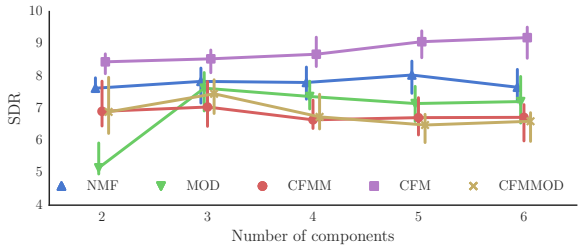

Fig. 3. Boxplots of SDR values of the unison dataset over the number of components j. For $j > 2$ oracle clustering was applied.

4. CONCLUSION

In this work we presented a method to exploit common modulation textures for source separation. A transformation based on a complex tensor representation computed from patches of the STFT has been introduced. We then showed how these patches are factorized by the proposed *Common Fate Model*, which is derived from the idea of humans perceiving common modulation over time as one source. Our results on unisonous musical instruments indicate that this method can perform well for this scenario. The CFM model could also be successfully used in other scenarios, such as speech separation.

[4] VIENNA SYMPHONIC LIBRARY (https://vsl.co.at)

[5] www.loria.fr/~aliutkus/cfm/

5. 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, no. 3, pp. 107–115, May 2014.

[2] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. WASPAA*, New Paltz, NY, USA, Oct. 2003, pp. 177–180.

[3] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," *Lecture Notes in Comp. Science*, vol. 3195, pp. 494–499, 2004.

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

[5] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in *Proc. ISSC*, Dublin, Ireland, 2005.

[6] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, pp. 15 pages, 2008, Article ID 872425.

[7] D. Fitzgerald and M. Cranitch, "Sound source separation using shifted non-negative tensor factorisation," in *Proc. ICASSP*, Toulouse, France, May 2006.

[8] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation : statistical insights and towards self-clustering of the spatial cues," in *Proc. CMMR*, Malaga, Spain, 2010, pp. 102–115.

[9] A. Ozerov, C. Févotte, R. Blouet, and J.-L. Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. ICASSP*, Prague, May 2011, pp. 257–260.

[10] F-R. Stöter, S. Bayer, and B. Edler, "Unison source separation," in *Proc. DAFx*, Erlangen, Germany, Sep. 2014.

[11] R. Hennequin, R. Badeau, and B. David, "NMF with time–frequency activations to model nonstationary audio events," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 744–753, 2011.

[12] S. Greenberg and B. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. ICASSP*, Munich, Germany, Apr. 1997, vol. 3, pp. 1647–1650.

[13] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[14] T. Kinnunen, K. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *Proc. Odyssey*, Stellenbosch, South Africa, 2008, p. 30.

[15] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *Proc. EMBC*, Minneapolis, USA, Sep. 2009, pp. 2514–2517.

[16] T. Barker and T. Virtanen, "Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation," in *Proc. Interspeech*, Lyon, France, 2013, pp. 827–831.

[17] R. Badeau, "Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. WASPAA*, New Paltz, NY, USA, Oct. 2011, pp. 253–256.

[18] R. Badeau and M.D. Plumbley, "Multichannel HR-NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain," in *Proc. WASPAA*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[19] R. Badeau and M.D. Plumbley, "Multichannel high resolution NMF for modelling convolutive mixtures of non-stationary signals in the time-frequency domain," *IEEE Trans. on Audio, Sp. & Lang. Proc.*, vol. 22, no. 11, pp. 1670–1680, Nov. 2014.

[20] P. Magron, R. Badeau, and B. David, "Phase recovery in NMF for audio source separation: an insightful benchmark," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 81–85.

[21] R. Badeau and A. Dremeau, "Variational Bayesian EM algorithm for modeling mixtures of non-stationary signals in the time-frequency domain (HR-NMF)," in *Proc. ICASSP*, Vancouver, BC, Canada, May 2013, pp. 6171–6175.

[22] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, A Bradford book. Bradford Books, 1994.

[23] 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, Mar. 2009.

[24] N.Q.K. Duong, E. Vincent, and R. Gribonval, "Underdetermined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 18, no. 7, pp. 1830 –1840, Sept. 2010.

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

[26] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Trans. on Signal Processing*, vol. 59, no. 7, pp. 3155 –3167, Jul. 2011.

[27] A. Liutkus and R. Badeau, "Generalized Wiener filtering with fractional power spectrograms," in *Proc. ICASSP*, Brisbane, Australia, April 2015.

[28] G. Samoradnitsky and M. Taqqu, *Stable non-Gaussian random processes: stochastic models with infinite variance*, vol. 1, CRC Press, 1994.

[29] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*, Wiley Publishing, Sept. 2009.

[30] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 66–75, May 2014.

[31] A. Liutkus, T. Olubanjo, E. Moore, and M. Ghovanloo, "Source separation for target enhancement of food intake acoustics from noisy recordings," in *Proc. WASPAA*, New Paltz, New York, USA, Oct. 2015.

[32] D. FitzGerald, M. Cranitch, and E. Coyle, "On the use of the beta divergence for musical source separation," in *Proc. ISSC*, Galway, Ireland, Jun. 2008.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.