Nonnegative tensor factorization with frequency modulation cues for blind audio source separation

Elliot Creager, Noah Stein, Roland Badeau, Philippe Depalle

NONNEGATIVE TENSOR FACTORIZATION WITH FREQUENCY MODULATION CUES FOR BLIND AUDIO SOURCE SEPARATION

Elliot Creager[1,3] **Noah D. Stein**[1] **Roland Badeau**[2,3] **Philippe Depalle**[3]
[1] Analog Devices Lyric Labs, Cambridge, MA, USA
[2] LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, Paris, France
[3] CIRMMT, McGill University, Montréal, Canada
elliot.creager@analog.com, noah.stein@analog.com,
roland.badeau@telecom-paristech.fr, depalle@music.mcgill.ca

ABSTRACT

We present Vibrato Nonnegative Tensor Factorization, an algorithm for single-channel unsupervised audio source separation with an application to separating instrumental or vocal sources with nonstationary pitch from music recordings. Our approach extends Nonnegative Matrix Factorization for audio modeling by including local estimates of frequency modulation as cues in the separation. This permits the modeling and unsupervised separation of vibrato or glissando musical sources, which is not possible with the basic matrix factorization formulation.

The algorithm factorizes a sparse nonnegative tensor comprising the audio spectrogram and local frequency-slope-to-frequency ratios, which are estimated at each time-frequency bin using the Distributed Derivative Method. The use of local frequency modulations as separation cues is motivated by the principle of common fate partial grouping from Auditory Scene Analysis, which hypothesizes that each latent source in a mixture is characterized perceptually by coherent frequency and amplitude modulations shared by its component partials. We derive multiplicative factor updates by Minorization-Maximization, which guarantees convergence to a local optimum by iteration. We then compare our method to the baseline on two separation tasks: one considers synthetic vibrato notes, while the other considers vibrato string instrument recordings.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [11] is a popular method for the analysis of audio spectrograms [16], especially for audio source separation [17]. NMF models the observed spectrogram as a weighted sum of rank-1 latent components, each of which factorizes as the outer product of a pair of vectors representing the constituent frequencies and onset regions for some significant component in the mixture, e.g. a musical note. Equivalently, the entire spectrogram matrix approximately factorizes as a matrix of spectral templates times a matrix of temporal activations, typically such that the approximate factors have many fewer elements than the full observation. While NMF can be used for supervised source separation tasks with a straightforward extension of the signal model [19], this necessitates pre-training NMF representations for each source of interest.

The use of modulation cues in source separation is popular in the Computational Auditory Scene Analysis (CASA) [26] literature, which, unlike NMF, typically relies on partial tracking. E.g., [25] isolates individual partials by frequency warping and filtering, while [12] groups partials via correlations in amplitude modulations. [2], which more closely resembles our work in the sense of being data-driven, factorizes a tensor encoding amplitude modulations for speech separation.

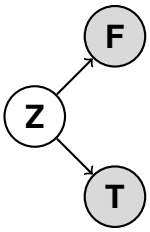
Our approach is inspired by [20] and [21], which present a Nonnegative Tensor Factorization (NTF) incorporating direction-of-arrival (DOA) estimates in an unsupervised speech source separation task. Whereas use of DOA information in that work necessitates multi-microphone data, we address the single-channel case by incorporating the local frequency modulation (FM) cues at each time-frequency bin. These cues are combined with the spectrogram as a sparse observation tensor, which we factorize in a probabilistic framework. The modulation cues are adopted structurally by way of an NTF where each source in the mixture is modeled via an NMF factor and a time-varying FM factor.

2. BACKGROUND

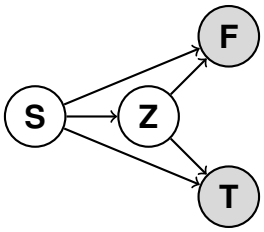
2.1 Nonnegative matrix factorization

We now summarize NMF within a probabilistic framework. We consider the normalized Short-Time Fourier Transform (STFT) magnitudes (i.e., spectrogram) of the input signal as an observed discrete probability distribution of energy over the time-frequency plane, i.e.,

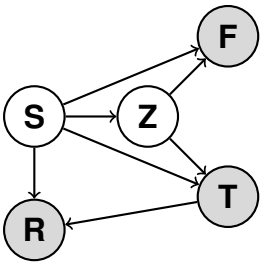
$$p^{\text{obs}}(f,t) \triangleq \frac{|X(f,t)|}{\sum_{\nu,\tau}|X(\nu,\tau)|}, \tag{1}$$



(a) Basic NMF (Section 2.1)



(b) NMF for source separation (Section 2.2)



(c) Vibrato NTF (Section 3.3)

Figure 1: Graphical models for the factorizations in this paper. In each case the input data are a distribution over the observed (shaded) variables, while the model approximates the observation by a joint distribution over observed and latent (unshaded) variables that factorizes as specified. F, T, Z, S, and R respectively represent the discrete frequencies, hops, components, sources, and frequency modulations over which the data is distributed.

$\forall f \in \{1, ..., F\}, t \in \{1, ..., T\}$, where X is the input STFT and (f, t) indexes the time-frequency plane. NMF seeks an approximation q to observed distribution p^{obs} that is a valid distribution over the time-frequency plane and factorizes as

$$q(f,t) = \sum_z q(f|z)q(t|z)q(z) = \sum_z q(f|z)q(z,t). \quad (2)$$

Figure 1(a) shows the graphical model for a joint distribution with this factorization.

We have introduced $z \in \{1, ..., Z\}$ as a latent variable that indexes components in the mixture, typically with Z chosen to yield an overall data reduction, i.e., $FZ + ZT \ll FT$. For a fixed z_0, $q(f|z_0)$ is a vector interpreted as the spectral template of the z_0-th component, i.e., the distribution over frequency bins of energy belonging to that component. Likewise, $q(z_0, t)$ is interpreted as a vector of temporal activations of the z_0-th component, i.e., it specifies at what time indices the z_0-th component is prominent in the observed mixture. Indeed, (2) can be implemented as a matrix multiplication, with the usual nonnegativity constraint on the factors satisfied implicitly, since q is a valid probability distribution.

The optimization problem is typically formalized as minimizing the Kullback-Leibler (KL) divergence between the observation and approximation, or equivalently as maximizing the cross entropy between the two distributions:

$$\begin{aligned} \underset{q}{\text{maximize}} \quad & \sum_{f,t} p^{\text{obs}}(f,t) \log q(f,t) \\ \text{subject to} \quad & q(f,t) = \sum_z q(f|z)q(z,t). \end{aligned} \quad (3)$$

While the non-convexity of this problem prohibits a globally optimal solution in reasonable time, a locally optimal solution can be found by multiplicative updates to the factors, which were first presented in [10]. We refer to this algorithm as KL-NMF, but note its equivalence to Probabilistic Latent Component Analysis (PLCA) [18], as well as a strong connection to topic modeling of counts data.

2.2 NMF for source separation

NMF can be leveraged as a source model within a source separation task, such that the observed mixture is modeled as a sum of sources, each of which is modeled by NMF. Whereas the latent variable z in NMF indexes latent components belonging to a source, we now introduce an additional latent variable $s \in \{1, .., S\}$, which indexes latent sources within the mixture. The resulting joint distribution over observed and latent variables is expressed as

$$q(f,t,s,z) = q(s)q(f|s,z)q(z,t|s). \quad (4)$$

Thus the approximation to $p^{obs}(f,t)$ is the marginal distribution

$$\begin{aligned} q(f,t) &= \sum_s q(s)q(f,t|s) \\ &= \sum_s q(s) \sum_z q(f|s,z)q(z,t|s), \end{aligned} \quad (5)$$

where $q(s_0)$ and $q(f,t|s_0)$ represent the mixing coefficient and NMF source model for the s_0-th source in the mixture, respectively. Figure 1(b) shows the graphical model.

Given a suitable approximation q, we estimate the latent sources in the mixture via Wiener filtering, i.e.,

$$X_s(f,t) = X(f,t)q(s|f,t), \quad (6)$$

where the Wiener gains $q(s|f,t)$ are given by the conditional probabilities [1] of the latent sources given the approximating joint distribution

$$q(s|f,t) = \frac{q(f,t,s)}{q(f,t)} = \frac{\sum_z q(s)q(f|s,z)q(z,t|s)}{\sum_{z,s'} q(s')q(f|s',z)q(z,t|s')}. \quad (7)$$

The estimated sources can then be reconstructed in the time-domain via the inverse STFT.

We seek a q that both approximates p^{obs} and yields source estimates $q(f,t|s)$ close to the true sources. In a supervised setting, the spectral templates for each source model can be fixed by using basic NMF on some characteristic training examples in isolation. When the appropriate training data is unavailable, the basic NMF can

[1] A convenient result of the Wiener filter gains being conditional distributions over sources is that the mixture energy is conserved by the source estimates in the sense that $\sum_s X_s(f,t) = X(f,t) \; \forall \, f, t$.

be extended by introducing priors on the factors or otherwise adding structure to the observation model to encourage, e.g., smoothness in the activations [24] or harmonicity in the spectral templates [3], which hopefully in turn improves the source estimates. By contrast, our approach exploits local FM cues directly in the factorization, yielding an observation model for latent sources consistent with the sorts of pitch modulations expected in musical sounds.

2.3 Coherent frequency modulation

We now introduce frequency-slope-to-frequency ratios (FSFR) as local signal parameters under an additive sinusoidal model that are useful as grouping cues for the separation of sources with coherent FM, e.g. in the vibrato or glissando effects. In continuous time, the additive sinusoidal model expresses the s-th source as a sum of component partials, [2] each parameterized by an instantaneous frequency and amplitude, i.e.,

$$x_s(\tau) = \sum_{p=1}^{P} A_p(\tau) \cos\left(\theta_p(\tau_0) + \int_{\tau_0}^{\tau} \omega_p(u) du\right) \quad (8)$$

where p is the partial index, and $\theta_p(\tau_0)$, $A_p(\tau)$ and $\omega_p(\tau)$ specify the initial phase, instantaneous amplitude, and instantaneous frequency of the p-th partial.

We now consider a source under coherent FM, i.e.,

$$\omega_p(\tau) \triangleq (1 + \kappa_s(\tau))\omega_p(\tau_0) \,\forall\, p \quad (9)$$

for some modulation function κ_s with $\kappa_s(\tau_0) = 0$. E.g., κ_s resembles a slowly-varying sinusoid during frequency vibrato, or a gradual ramp function during glissando. The FSFR are then expressed as

$$\upsilon_p(\tau) \triangleq \frac{\omega_p'(\tau)}{\omega_p(\tau)} = \frac{\kappa_s'(\tau)}{1 + \kappa_s(\tau)}. \quad (10)$$

Note that $\{\upsilon_p(\tau)\}$ are time-varying but independent of the partial index p for a given source index s. In other words, the instantaneous FSFR is common to all partials belonging to the same source and can be used as a grouping cue in unsupervised source separation [7].

2.4 Distributed Derivative Method

We now summarize the Distributed Derivative Method (DDM) [4, 8] for signal parameter estimation, which we use to estimate the FSFR at each time-frequency bin. DDM estimates the parameters of a monochrome analytic signal under a Q-th order generalized sinusoid model, [3] which is expressed as

$$x(\tau) = \exp\left(\sum_{q=0}^{Q} \eta_q \tau^q\right), \quad (11)$$

where $\boldsymbol{\eta} \in \mathbb{C}^{Q+1}$ is the vector of signal parameters, whose real and imaginary parts specify the log amplitude law and phase law, [4] respectively. In this work, we specify (11) as a constant amplitude signal with linear frequency modulation, i.e., $\boldsymbol{\eta} \in \mathbb{C}^3$ with $\Re(\eta_i) = 0 \;\; \forall \;\; i$. The signal parameters $\Im(\eta_1)$ and $\Im(\eta_2)$ then specify (within multiplicative constants) the instantaneous frequency and frequency slope, respectively.

The parameters of interest can be estimated by considering the inner product of the signal with a family of differentiable analysis atoms of finite time-frequency support. In particular, the continuous-time STFT can be expressed by inner product as

$$\mathcal{X}(f,t) \triangleq \langle x(\tau), \phi(\tau; f, t)\rangle = \int_{\tau=-\infty}^{+\infty} x(\tau)\phi(\tau; f, t)^* d\tau, \quad (12)$$

where $\mathcal{X}(f,t)$ is the STFT, $x(\tau)$ is the input signal, and $\phi(\tau; f, t)$ is a heterodyned window function from some differentiable family (e.g. Hann), parameterized by its localization (f, t) in the time-frequency plane. The signal parameters are solutions to equations of the form

$$\langle x(\tau), \phi'(\tau; f, t)\rangle = -\sum_{q=1}^{Q} \eta_q \langle q\tau^{q-1} x(\tau), \phi(\tau; f, t)\rangle, \quad (13)$$

which is linear in $\{\eta_q\}$ for $q > 0$, and permits an STFT-like computation of both inner products. The right-hand side of (13) is derived from the left-hand side using integration by parts, exploiting the finite support of $\phi(\tau; f, t)$, and substituting in the signal derivative $x'(\tau)$ from (11). To estimate the signal parameters at a particular (f_0, t_0), we construct a system of linear equations by evaluating (13) for each $\phi(\tau; f, t)$ in a set of nearby atoms Φ, then solve for $\boldsymbol{\eta}$ in a least-squares sense. We typically use atoms in neighboring frequency bins at the same time step, i.e., $\Phi = \{\phi(\tau; t_0, f_0 - \frac{L-1}{2}),\ ...,\ \phi(\tau; t_0, f_0 + \frac{L-1}{2})\}$ for some odd L.

While DDM is an unbiased estimator of the signal parameters in continuous time, we must implement a discrete-time approximation on a computer. This introduces a small bias that can be ignored in practice since the STFT window is typically longer than a few samples [4].

3. PROPOSED METHOD

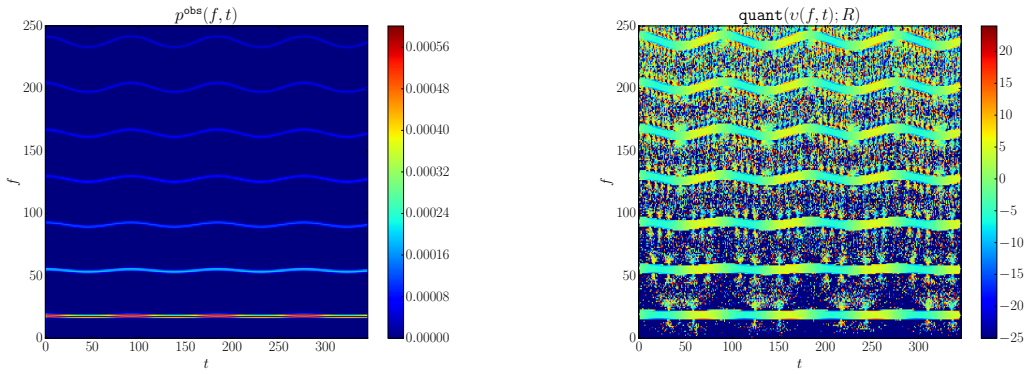
3.1 Motivation

The NMF signal model is not sufficiently expressive to compactly represent a large class of musical sounds, namely those characterized by slow frequency modulations, e.g., in the vibrato effect. In particular, it specifies a single fixed spectral template per latent component

[2] We do not assume any special structure in the partial frequencies, e.g., harmonicity.

[3] It is natural to specify the signal locally (near some time-frequency bin) as a generalized sinusoid even while the global model remains additive sinusoidal. In particular, the notion of a time-frequency-localized signal follows from the filterbank summation interpretation of the STFT, and corresponds to the heterodyned and shifted input, prior to low-pass filtering by the window and downsampling in time [1]. In a slight abuse of notation, we later absorb the time-frequency indices as parameters in the analysis atom, i.e., we switch to the overlap-add interpretation of the STFT without warning.

[4] The frequency law is trivially computed from the phase law.



(a) Spectrogram

(b) Discretized FSFR

Figure 2: Unfolding the nonzero elements in the observation tensor for a synthetic vibrato square wave note (G5). The hop index t spans 2 seconds of the input audio, while the bin index f spans half the sampling rate, 0–22.05 kHz.

and thus requires a large number of components to model sounds with nonstationary pitch. From a separation perspective, as the number of latent components grows, so grows the need for a comprehensive model that can correctly group components belonging to the same source. To this end, we appeal to the perceptual theory of Auditory Scene Analysis [5], which postulates the importance of shared frequency or amplitude modulations among partials as a perceptual cue in their grouping [6, 14]. In this work we focus on FM, although in principle our approach could be extended to include amplitude modulations. [5] We now propose an extension to KL-NMF that leverages this so-called common fate principle and is suitable for the analysis of vibrato signals.

3.2 Compiling the observations as a tensor

DDM yields the local estimates of frequency and frequency slope for each time-frequency bin, from which the FSFR are trivially computed. We define the (sparse) observation tensor $p^{\text{obs}}(f,t,r) \in \mathbb{R}_{\geq 0}^{F\times T\times R}$ as an assignment of the normalized spectrogram into one of R discrete bins for each (f,t) according the local FSFR estimate, i.e.,

$$p^{\text{obs}}(f,t,r) \triangleq \begin{cases} p^{\text{obs}}(f,t) & \text{if} \quad \texttt{quant}(v(f,t);R) = r \\ 0 & \text{else,} \end{cases} \tag{14}$$

where $p^{\text{obs}}(f,t)$ is the normalized spectrogram as in (1) and v are the FSFR as in (10), which are quantized by $\texttt{quant}(\cdot;R)$, possibly after clipping to some reasonable range of values. Figure 2 shows the spectrogram and FSFR for a synthetic vibrato square wave.

3.3 Vibrato NTF

As with NMF, we seek a joint distribution q with a particular factorized form, whose marginal maximizes cross entropy against the observed data. We propose an observation model of the form

$$q(f,t,r) = \sum_s q(s)q(r|t,s)\sum_z q(f|s,z)q(z,t|s) \tag{15}$$

where $q(s)$ represents the mixing, $q(r|t,s)$ represents the common time-varying FSFR per source, and $\sum_z q(f|s,z)q(z,t|s)$ represents the NMF source model. Figure 1(c) shows the graphical model of the joint distribution. Thus, given p^{obs}, we seek an approximation q that factorizes as in (15) and maximizes

$$\begin{aligned} \alpha(q) &\triangleq \sum_{f,t,r} p^{\text{obs}}(f,t,r)\log q(f,t,r) \\ &= \sum_{f,t,r} p^{\text{obs}}(f,t,r)\log\sum_{z,s} q(f,t,r,z,s). \end{aligned} \tag{16}$$

The sum in the argument to the log makes this difficult to solve outright, so we find a local optimum by iterative Minorization-Maximization (MM) [9] instead. That is, given $q^{(i)}$, our model at the current (i-th) iteration, we pick a better $q^{(i+1)}$ by (a) finding a concave minorizing function $\beta(q;q^{(i)})$ such that $\beta(q;q^{(i)}) \leq \alpha(q)\ \forall\ q$ and $\beta(q^{(i)};q^{(i)}) = \alpha(q^{(i)})$, and (b) maximizing $\beta(q;q^{(i)})$ with respect to q.

In particular, $\beta(q;q^{(i)})$ is derived [6] by applying Jensen's inequality to (16), and is expressed as

$$\beta(q;q^{(i)}) \triangleq \sum_{f,t,r,z,s} p^{\text{obs}}(f,t,r)q^{(i)}(z,s|f,t,r)\log\frac{q(f,t,r,z,s)}{q^{(i)}(z,s|f,t,r)}, \tag{17}$$

where $q^{(i)}(z,s|f,t,r)$ is the approximate posterior over latent variables given the model at the i-th iteration [7], computed as

$$q^{(i)}(z,s|f,t,r) = \frac{q^{(i)}(z,s,f,t,r)}{\sum_{z',s'} q^{(i)}(z',s',f,t,r)}. \tag{18}$$

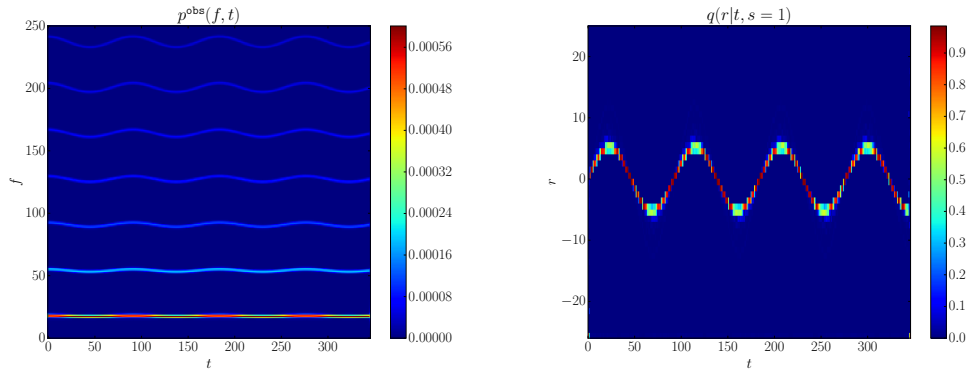
For notational convenience we define $\rho(f,t,r,z,s) \triangleq p^{\text{obs}}(f,t,r)q^{(i)}(z,s|f,t,r)$ and discarding the denominator in the log of (17) (constant w.r.t. q), equivalently write the optimization over the minorizing function as

$$\max_q \sum_{f,t,r,z,s} \rho(f,t,r,z,s)\log q(s)q(f|z,s)q(z,t|s)q(r|t,s). \tag{19}$$

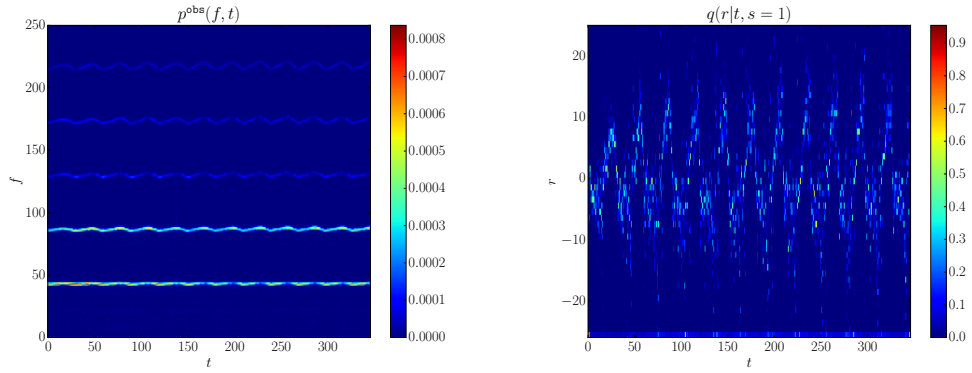
[5] In turn, this would increase the dimensionality of the data.

[6] Cf. [20] for a more thorough treatment.

[7] Note that the MM iteration specifies an expectation-maximization.



(a) Spectrogram (b) VibNTF FM model



(c) Spectrogram (d) VibNTF FM model

Figure 3: For single-note analyses, VibNTF encodes the time-varying pitch modulation. The top row shows a synthetic vibrato square wave note (G5), while the bottom row shows a real recording of a violin vibrato note (B♭6). We plot r in the range $[-\frac{R}{2}, \frac{R}{2}]$ in figures 3(b) and 3(d) to clarify that the index r represents a zero-mean quantity (the FSFR).

We now alternatively update each factor by separating the argument in the log in (19) as a sum of logs, each term of which can be optimized by applying Gibb's inequality [13]. That is, given the current model, the optimal choice for some factor of $q^{(i+1)}$ is the marginal of ρ over the corresponding variables. E.g.,

$$q^{(i+1)}(s) \leftarrow \frac{\sum_{f,t,r,z} \rho(f,t,r,z,s)}{\sum_{f,t,r,z,s'} \rho(f,t,r,z,s')}. \tag{20a}$$

Likewise, the remaining factor updates are expressed as

$$q^{(i+1)}(f|z,s) \leftarrow \frac{\sum_{t,r} \rho(f,t,r,z,s)}{\sum_{f',t,r} \rho(f',t,r,z,s)}; \tag{20b}$$

$$q^{(i+1)}(z,t|s) \leftarrow \frac{\sum_{f,r} \rho(f,t,r,z,s)}{\sum_{f,t',r,z'} \rho(f,t',r,z',s)}; \tag{20c}$$

$$q^{(i+1)}(r|t,s) \leftarrow \frac{\sum_{f,z} \rho(f,t,r,z,s)}{\sum_{f,r',z} \rho(f,t,r',z,s)}. \tag{20d}$$

Since ρ is expressed as a product of the current factors and observed data, the factor updates can be implemented efficiently by using matrix multiplications to sum across inner dimensions as necessary. The theory guarantees convergence [8] to a local minimum [9], although in practice we stop the algorithm after some fixed number of iterations. The algorithm is initialized by choosing factors of $q^{(0)}$ as random valid conditional probabilities.

Figure 3 visualizes the FM factor $q(r|t,s)$ estimated by the proposed algorithm for single note analyses ($S = 1$) of both synthetic and real data.

4. EVALUATION

We present a comparison of our proposed method with the baseline KL-NMF (which our method extends) in a blind source separation task examining mixtures of two single-note recordings. We use the `BSS_EVAL` criteria [23] to evaluate separation performance, which necessitates the use of artificial mixtures. We report the source-to-distortion ratio (SDR), source-to-interference ratio (SIR), and source-to-artifact ratio (SAR), each in dB. Each experiment comprises 500 separations, with the sources in each trial chosen as specified below and mixed at 0 dB with a total mixture duration of 2 seconds at 44.1 kHz sampling rate. We report the average metrics across all sources and trials.

To use KL-NMF for blind source separation, we must specify $Z = 2$, i.e., each mixture component considered as a source. This baseline should be relatively easy to beat, since empirically KL-NMF does a poor job of modeling vibrato signals when Z is small.

[8] For guaranteed convergence, ρ must be recomputed after each factor update, rather than once per iteration as the notation suggests. However, in practice we observe convergence without the recomputation.

	BSS_EVAL in dB		
Algorithm	SDR	SIR	SAR
(A) Synthetic data			
2-part KL-NMF	-1.5 $\pm$ 0.1	0.1 $\pm$ 0.2	6.9 $\pm$ 0.2
Vibrato NTF	14.6 $\pm$ 1.0	17.0 $\pm$ 1.2	23.6 $\pm$ 0.7
(B) Real data			
2-part KL-NMF	2.8 $\pm$ 0.4	8.0 $\pm$ 2.1	9.2 $\pm$ 0.2
Vibrato NTF	5.8 $\pm$ 0.5	9.7 $\pm$ 2.2	17.7 $\pm$ 0.5

Table 1: Mean and 95% confidence intervals of the BSS_EVAL metrics for 500 unsupervised separations of two-source mixtures. Experiment A considers synthetic vibrato square waves, while experiment B considers single-note vibrato string instrument recordings.

For Vibrato NTF, we specify $S = 2$ and $Z = 3$, i.e., for each of the two sources we learn spectral templates and temporal activations for three components. E.g., considering a sinusoidal vibrato, the components could model the source during the crest, midpoint, and trough of the pitch modulation. We estimate the signal parameters at a particular (f_0, t_0) using DDM with a family of $L = 5$ analysis atoms (heterodyned Hann functions) in the same hop index and nearby frequency bins. In order to avoid the influence of noisy FSFR estimates in the factorization, we apply some mild post-processing prior to quantization. Specifically, we implicitly discard FSFR at (f, t) with $p^{\text{obs}}(f, t)$ below the 10^{th} percentile, or outside a reasonable range of ± 4 times the sampling rate by setting them to the data median. The FSFR are then quantized evenly across their range into $R = 50$ discrete values.

For both algorithms, the STFT in (1) is specified by a 1024-length (23 msec) Discrete Fourier Transform using a Hann window with 75% overlap between successive frames. Thus, $F = 513$, corresponding to the non-redundant frequency bins, and $T = 346$, the number of hops required to cover the mixture duration. Both algorithms are initialized randomly and run for 100 iterations.

Experiment A examines synthetic data, where the sources are square waves with frequency vibrato, whose signal parameters are generated at random. The fundamental frequency corresponds to a note value selected uniformly at random from the three-octave range [A3, G$\sharp$5]. The number of partials is chosen uniformly at random from the range [10, 30], and subsequently reduced as necessary to avoid aliasing. The vibrato modulation function, i.e., κ_s in (9), is a sinusoid with depth chosen uniformly at random in the range of [5%, 20%] of the fundamental and rate chosen log-uniformly at random from the range [0.5, 10] Hz.

Experiment B examines real data, where the sources are single-note recordings from the McGill University Master Samples (MUMS) [15], which contains over 6000 single-note and single-phrase recordings of classical and popular instruments. We focus our evaluation on string instruments, which exhibit strong frequency modulation in their vibrato effect [22]. The MUMS subset of string instrument notes with vibrato comprises a total of 250 unique recordings of violin, viola, cello, and double bass. The sources are chosen randomly from this subset and trimmed or padded to 2 seconds as necessary.

Results for both experiments are provided in table 1. Experiment A shows a dramatic win for Vibrato NTF over the baseline. We see some variability in the results, which reflects an optimization over a cost surface with many local optima. With random initialization, Vibrato NTF works either very well or very poorly, so robustness could be improved by a more careful initialization, or alternatively by regularizing the factorization in such a way as to avoid sub-optimal solutions.

In experiment B, we see that moving from synthetic to real data degrades the performance of our proposed method, although we still beat the baseline by a modest margin. Interestingly, the baseline performs better on real data than synthetic, likely because the pitch variations are less pronounced so KL-NMF fails less frequently. Moreover, the pitch modulations in real data are more complex than in the synthetic case (compare figures 3(b) and 3(d)), and may require more components (larger Z) to be properly modeled. Vibrato NTF as proposed tends to decrease in performance as Z increases, so additional work is required to improve robustness for the analysis of real data. We hypothesize that an extension enforcing temporal continuity in the FM factor, which should be smooth and monotonic per-source, would enhance the grouping of components, permitting a larger Z in practice.

5. CONCLUSION

We proposed Vibrato NTF, a novel blind source separation algorithm that extends NMF by leveraging local estimates of frequency modulation as grouping cues directly in the factorization. Experimental results using synthetic data showed a substantial improvement over the baseline, and validated the FSFR as useful grouping cues in a source separation task. In the experiment with real recordings, our method provided a more modest improvement. With regards to the analysis of real data, we believe the incorporation of sensible priors on the factors would improve the separation performance, while careful initalization would improve the robustness. Further work could include tailoring the proposed method to the analysis of polyphonic sounds, or sounds with mild or no frequency modulation. Additionally, an extension including coherent amplitude modulations as a grouping cue is possible within the proposed tensor factorization framework.

6. ACKNOWLEDGEMENTS

The research leading to this paper was partially supported by the French National Research Agency (ANR) as a part of the EDISON 3D project (ANR- 13-CORD-0008-02), and by the Canadian National Science and Engineering Research Council (NSERC). Additional support was provided by the Analog Garage, the emerging business accelerator at Analog Devices, Inc.

7. REFERENCES

[1] J. Allen and L. Rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE*, 65:1558–64, 1977.

[2] T. Barker and T. Virtanen. Non-negative tensor factorization of modulation spectrograms for monaural sound separation. In *Proceedings of the 2013 Interspeech Conference*, pages 827–31, Lyon, France, 2013.

[3] 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*, 18(3):538–49, 2010.

[4] M. Betser. Sinusoidal polyphonic parameter estimation using the distribution derivative. *IEEE Transactions on Signal Processing*, 57(12):4633–45, 2009.

[5] A. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, Cambridge, MA, 1990.

[6] J. M. Chowning. Computer synthesis of the singing voice. In *Sound Generation in Winds, Strings, Computers*, pages 4–13. Kungl. Musikaliska Akademien, Stokholm, Sweden, 1980.

[7] E. Creager. Musical source separation by coherent frequency modulation cues. Master's thesis, McGill University, 2015.

[8] B. Hamilton and P. Depalle. A unified view of non-stationary sinusoidal parameter estimation methods using signal derivatives. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 369–72, Kyoto, Japan, 2012.

[9] D. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–7, 2004.

[10] D. Lee, M. Hill, and H. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–62, 2001.

[11] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–91, 1999.

[12] Y. Li, J. Woodruff, and D.Wang. Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(7):1361–71, 2009.

[13] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, UK, 2005.

[14] S. McAdams. Segregation of concurrent sounds I: Effects of frequency modulation coherence. *Journal of the Acoustic Society of America*, 86(6):2148–59, 1989.

[15] F. Opolko and J. Wapnick. McGill University master samples [Compact Disks], 1987.

[16] P. Smaragdis and J. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–80, New Paltz, NY, 2003.

[17] P. Smaragdis, C. Févotte, G. Mysore, N. Mohammadia, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–74, 2014.

[18] P. Smaragdis, B. Raj, and M. Shashanka. A probabilistic latent variable model for acoustic modeling. In *Proceedings of the NIPS Workshop of Advances in Models for Acoustic Processing*, Vancouver, Canada, 2006.

[19] P. Smaragdis, B. Raj, and M. Shashanka. Supervised and semi-supervised separation of sounds single-channel mixtures. *Independent Component Analysis and Signal Separation*, (Lecture Notes in Computer Science, 4666):414–21, 2007.

[20] N. Stein. Nonnegative tensor factorization for directional unsupervised audio source separation. *arXiv preprint*, http://arxiv.org/abs/1411.5010, 2015.

[21] J. Traa, P. Smaragdis, N. Stein, and D. Wingate. Directional NMF for joint source localization and separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2015.

[22] V. Verfaille, C. Guastavino, and P. Depalle. Perceptual evaluation of vibrato models. In *Proceedings of the Conference on Interdisciplinary Musicology*, Montreal, Canada, 2005.

[23] E. Vincent, R. Gribonval, and C. Févotte. Performance measurements in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–9, 2006.

[24] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–74, 2007.

[25] A. Wang. Instantaneous and frequency-warped techniques for source separation and signal parameterization. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 47–50, New Paltz, NY, 1995.

[26] D. Wang and G. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley Interscience, Hoboken, NJ, 2006.