# EXPLOITING LONG-TERM TEMPORAL DEPENDENCIES IN NMF USING RECURRENT NEURAL NETWORKS WITH APPLICATION TO SOURCE SEPARATION

*Nicolas Boulanger-Lewandowski*[*]

Université de Montréal
Montréal, QC, Canada

*Gautham J. Mysore* *Matthew Hoffman*

Adobe Research
San Francisco, CA, USA

## ABSTRACT

This paper seeks to exploit high-level temporal information during feature extraction from audio signals via non-negative matrix factorization. Contrary to existing approaches that impose local temporal constraints, we train powerful recurrent neural network models to capture long-term temporal dependencies and event co-occurrence in the data. This gives our method the ability to "fill in the blanks" in a smart way during feature extraction from complex audio mixtures, an ability very useful for a number of audio applications. We apply these ideas to source separation problems.

***Index Terms***— Recurrent neural networks, long-term temporal dependencies, non-negative matrix factorization, audio source separation

## 1. INTRODUCTION

Non-negative matrix factorization (NMF) is an unsupervised technique to discover parts-based representations underlying non-negative data [1]. When applied to the magnitude spectrogram of an audio signal, NMF can discover a basis of interpretable recurring events and their associated time-varying encodings, or *activities*, that together optimally reconstruct the original spectrogram. In addition to accurate reconstruction, it is often useful to enforce various constraints to influence the decomposition. Those constraints generally act on each time frame independently to encourage sparsity [2], harmonicity of the basis spectra [3] or relevance with respect to a discriminative criterion [4], or include a temporal component such as simple continuity [5, 6, 7, 8], Kalman filtering like techniques [9, 10, 11] or Markov chain modeling [12, 13, 14, 15]. In this paper, we aim to improve the temporal description in the latter category with an expressive connectionist model that can describe long-term dependencies and high-level structure in the data.

Recurrent neural networks (RNN) [16] are powerful dynamical systems that incorporate an internal memory, or *hidden state*, represented by a self-connected layer of neurons. This property makes them well suited to model temporal sequences, such as frames in a magnitude spectrogram or feature vectors in an activity matrix, by being trained to predict the output at the next time step given the previous ones. RNNs are completely general in that in principle they can describe arbitrarily complex long-term temporal dependencies, which has made them very successful in music, language and speech applications [17, 18, 19, 20]. A recent extension of the RNN, called the RNN-RBM, employs time-dependent restricted Boltzmann machines (RBM) to describe the multimodal conditional densities typically present in audio signals, resulting in significant improvements over N-gram and HMM baselines [21, 17]. In this paper, we show how to integrate RNNs into the NMF framework in order to model sound mixtures. We apply our approach to audio source separation problems, but the technique is general and can be used for various audio applications.

The remainder of the paper is organized as follows. In sections 2 and 3 we introduce the NMF and RNN models. In section 4 we incorporate temporal constraints into the feature extraction algorithm. Finally, we present our methodology and results in sections 5 and 6.

## 2. NON-NEGATIVE MATRIX FACTORIZATION

The NMF method aims to discover an approximate factorization of an input matrix $X$:

$$\overset{N\times T}{X} \simeq \overset{N\times T}{\Lambda} \equiv \overset{N\times K}{W} \cdot \overset{K\times T}{H}, \tag{1}$$

where $X$ is the observed magnitude spectrogram with time and frequency dimensions $T$ and $N$ respectively, $\Lambda$ is the reconstructed spectrogram, $W$ is a dictionary matrix of $K$ basis spectra and $H$ is the activity matrix. Non-negativity constraints $W_{nk} \geq 0, H_{kt} \geq 0$ apply on both matrices. NMF seeks to minimize the *reconstruction error*, a distortion measure between the observed spectrogram $X$ and the reconstruction $\Lambda$. A popular choice is the generalized Kullback-Leibler divergence:

$$C_{KL} \equiv \sum_{nt} \left( X_{nt} \log \frac{X_{nt}}{\Lambda_{nt}} - X_{nt} + \Lambda_{nt} \right), \tag{2}$$

[*]This investigation was carried out during his internship at Adobe Research.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

with which we will demonstrate our method. Minimizing $C_{KL}$ can be achieved by alternating multiplicative updates to $H$ and $W$ [22]:

$$H \leftarrow H \circ \frac{W^T(X/\Lambda)}{W^T 1 1^T} \tag{3}$$

$$W \leftarrow W \circ \frac{(X/\Lambda)H^T}{1 1^T H^T}, \tag{4}$$

where 1 is a vector of ones, the $\circ$ operator denotes element-wise multiplication, and division is also element-wise. These updates are guaranteed to converge to a stationary point of the reconstruction error.

It is often reasonable to assume that active elements $H_{kt}$ should be limited to a small subset of the available basis spectra. To encourage this behavior, a sparsity penalty $C_S \equiv \lambda|H|$ can be added to the total NMF objective [23], where $|\cdot|$ denotes the $L_1$ norm and $\lambda$ specifies the relative importance of sparsity. In that context, we impose the constraint that the basis spectra have unit norm. Equation (3) becomes:

$$H \leftarrow H \circ \frac{W^T(X/\Lambda)}{1+\lambda}, \tag{5}$$

and the multiplicative update to $W$ (eq. 4) is replaced by projected gradient descent [24]:

$$W \leftarrow W - \mu(1 - X/\Lambda)H^T \tag{6}$$

$$W_{nk} \leftarrow \max(W_{nk}, 0), W_{:k} \leftarrow \frac{W_{:k}}{|W_{:k}|}, \tag{7}$$

where $W_{:k}$ is the $k$-th column of $W$ and $\mu$ is the learning rate.

## 3. RECURRENT NEURAL NETWORKS

The RNN formally defines the distribution of the vector sequence $v \equiv \{v^{(t)} \in \mathbb{R}_0^{+K}, 1 \le t \le T\}$ of length $T$:

$$P(v) = \prod_{t=1}^{T} P(v^{(t)}|\mathcal{A}^{(t)}), \tag{8}$$

where $\mathcal{A}^{(t)} \equiv \{v^{(\tau)}|\tau < t\}$ is the sequence history at time $t$, and $P(v^{(t)}|\mathcal{A}^{(t)})$ is the conditional probability of observing $v^{(t)}$ according to the model, defined below.

A single-layer RNN with hidden units $\hat{h}^{(t)}$ is defined by its recurrence relation:

$$\hat{h}^{(t)} = \sigma(W_{v\hat{h}} v^{(t)} + W_{\hat{h}\hat{h}} \hat{h}^{(t-1)} + b_{\hat{h}}), \tag{9}$$

where $\sigma(x) \equiv (1 + e^{-x})^{-1}$ is the element-wise logistic sigmoid function, $W_{xy}$ is the weight matrix tying vectors $x, y$ and $b_x$ is the bias vector associated with $x$.

The model is trained to predict the observation $v^{(t)}$ at time step $t$ given the previous ones $\mathcal{A}^{(t)}$. The prediction $y^{(t)}$ is obtained from the hidden units at the previous time step $\hat{h}^{(t-1)}$:

$$y^{(t)} = o(W_{\hat{h}v} \hat{h}^{(t-1)} + b_v), \tag{10}$$

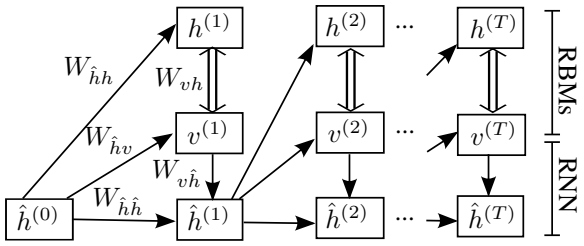



**Fig. 1**. Graphical structure of the RNN-RBM. Single arrows represent a deterministic function, double arrows represent the stochastic hidden-visible connections of an RBM.

where $o(a)$ is the output non-linearity function of an activation vector $a$, and should be as close as possible to the target vector $v^{(t)}$. When the target is a non-negative real-valued vector, the likelihood of an observation can be given by:

$$P(v^{(t)}|\mathcal{A}^{(t)}) \propto \frac{v^{(t)} \cdot y^{(t)}}{|v^{(t)}| \cdot |y^{(t)}|} \tag{11}$$

$$o(a)_k = \exp(a_k). \tag{12}$$

Other forms for $P$ and $o$ are possible; we have found that the cosine distance combined with an exponential non-linearity work well in practice, presumably because predicting the *orientation* of a vector is much easier for an RNN than predicting its magnitude.

When the output observations are multivariate, another approach is to capture the higher-order dependencies between the output variables using a powerful output probability model such as an RBM, resulting in the so-called RNN-RBM (Figure 1) [21, 17]. The Gaussian RBM variant is typically used to estimate the density of real-valued variables $v^{(t)}$ [25]. In this case, the RNN's task is to predict the parameters of the conditional distribution, i.e. the RBM biases at time step $t$:

$$b_v^{(t)} = b_v + W_{\hat{h}v} \hat{h}^{(t-1)} \tag{13}$$

$$b_h^{(t)} = b_h + W_{\hat{h}h} \hat{h}^{(t-1)}. \tag{14}$$

In an RBM, the likelihood of an observation is related to the free energy $F(v^{(t)})$ by $P(v^{(t)}|\mathcal{A}^{(t)}) \propto e^{-F(v^{(t)})}$:

$$F(v^{(t)}) \equiv \frac{1}{2}||v^{(t)}||^2 - b_v^{(t)} \cdot v^{(t)} - |s(b_h^{(t)} + W_{vh} v^{(t)})|, \tag{15}$$

where $s(x) \equiv \log(1 + e^x)$ is the element-wise softplus function and $W_{vh}$ is the weight matrix of the RBM. The log-likelihood gradient with respect to the RBM parameters is generally intractable due to the normalization constant but can be estimated by contrastive divergence [26, 17].

The RNN model can be trained by minimizing the negative log-likelihood of the data:

$$C_{RNN}(v) = -\sum_{t=1}^{T} \log P(v^{(t)}|\mathcal{A}^{(t)}), \tag{16}$$

whose gradient with respect to the RNN parameters is obtained by backpropagation through time (BPTT) [16]. Several strategies can be used to reduce the difficulties associated with gradient-based learning in RNNs including gradient clipping, sparsity and momentum techniques [27, 20].

## 4. TEMPORALLY CONSTRAINED NMF

In this section, we incorporate RNN regularization into the NMF framework to temporally constrain the activity matrix $H$ during the decomposition. A simple form of regularization that encourages neighboring activity coefficients to be close to each other is temporal smoothing:

$$C_{TS} = \frac{1}{2}\beta \sum_{t=1}^{T-1} ||H_{:t} - H_{:t+1}||^2, \tag{17}$$

where the hyperparameter $\beta$ is a weighting coefficient.

In the proposed model, we add the RNN negative log-likelihood term (eq. 16) with $v := \{H_{:t}, 1 \leq t \leq T\}$ to the total NMF cost:

$$C = C_{KL} + C_S + C_{TS} + C_{L2} + \alpha C_{RNN}(H), \tag{18}$$

where $C_{L2} \equiv \frac{1}{2}\eta||H||^2$ provides $L_2$ regularization, and the hyperparameters $\eta, \alpha$ specify the relative importance of each prior. This framework corresponds to an RNN generative model at temperature $\alpha^{-1}$ describing the evolution of the latent variable $H_{:t}$, the observation $X_{:t}$ at time $t$ being conditioned on $H_{:t}$ via the reconstruction error $C_{KL}$. The overall graphical model can be seen as a generalization of the non-negative hidden Markov model (N-HMM) [15].

The NMF model is first trained in the usual way by alternating the updates (5)–(7) and extracting the activity features $H$; the RNN is then trained to minimize $C_{RNN}(H)$ by stochastic gradient descent. During supervised NMF [28], it is necessary to infer the activity matrix $H$ that minimizes the total cost (eq. 18) given a pre-trained dictionary $W$ and a test observation $X$. Our approach is to replace the multiplicative udpate (5) with a gradient descent update:

$$H \leftarrow H - \mu \left[W^T(1 - X/\Lambda) + \lambda + \eta H + \tfrac{\partial C_{TS}}{\partial H} + \alpha \tfrac{\partial C_{RNN}}{\partial H}\right] \tag{19}$$

where the gradient of $C_{TS}$ is given by:

$$\frac{\partial C_{TS}}{\partial H_{kt}} = \beta \begin{cases} H_{kt} - H_{k(t+1)} & \text{if } t = 1 \\ 2H_{kt} - H_{k(t-1)} - H_{k(t+1)} & \text{if } 1 < t < T \\ H_{kt} - H_{k(t-1)} & \text{if } t = T. \end{cases} \tag{20}$$

When deriving $\partial C_{RNN}/\partial H$, it is important to note that $H_{:t}$ affects the cost directly by matching the prediction $y^{(t)}$ in equation (11), and also indirectly by influencing the future predictions of the RNN via $\mathcal{A}^{(t+\delta t)}$. By fully backpropagating the gradient through time, we effectively take into account future observations $X_{:(t+\delta t)}$ when updating $H_{:t}$. While other existing approaches require sophisticated inference procedures [29, 30], the search for a *globally* optimal $H$ can be facilitated by using gradient descent when the inferred variables are real-valued.

The RNN-RBM requires a different approach due to the intractable partition function of the $t^{\text{th}}$ RBM that varies with $\mathcal{A}^{(t)}$. The retained strategy is to consider $\mathcal{A}^{(t)}$ fixed during inference and to approximate the gradient of the cost by:

$$\frac{C_{RNN}}{\partial v^{(t)}} \simeq \frac{\partial F(v^{(t)})}{\partial v^{(t)}} = v^{(t)} - b_v^{(t)} - \sigma(b_h^{(t)} + W_{vh}v^{(t)})W_{vh}^T. \tag{21}$$

Since this approach can be unstable, we only update the value of $\mathcal{A}^{(t)}$ every $m$ iterations of gradient descent ($m = 10$) and we use an RNN in conjunction with the RNN-RBM to exploit its tractability and norm independence properties.

## 5. EVALUATION

In the next section, we evaluate the performance of our RNN model on a source separation task in comparison with a traditional NMF baseline and NMF with temporal smoothing. Source separation is interesting for our architecture because, contrary to purely discriminative tasks such as multiple pitch estimation or chord estimation where RNNs are known to outperform other models [29, 30], source separation requires accurate signal reconstruction.

We consider the supervised and semi-supervised NMF algorithms [28] that consist in training submodels on isolated sources before concatenating the pre-trained dictionaries and feeding the relevant activity coefficients into the associated temporal model; final source estimates are obtained by separately reconstructing the part of the observation explained by each submodel. In the semi-supervised setting, an additional dictionary is trained from scratch for each new examined sequence and no temporal model is used for the unsupervised channel. Wiener filtering is used as a final step to ensure that the estimated source spectrograms $X^{(i)}$ add up to the original mixture $X$:

$$\hat{X}^{(i)} = \frac{X^{(i)}}{\sum_j X^{(j)}} \circ X, \tag{22}$$

before transforming each source in the time domain via the inverse short-term Fourier transform (STFT).

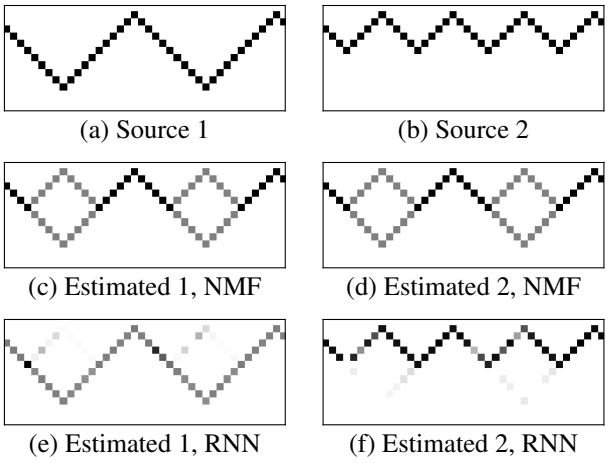
Our main experiments are carried out on the MIR-1K dataset[1] featuring 19 singers performing a total of 1,000 Chinese pop karaoke song excerpts, ranging from 4 to 13 seconds and recorded at 16 kHz. For each singer, the available tracks are randomly split into training, validation and test sets in a 8:1:1 ratio. The accompaniment music and singing voice channels are summed directly at their original loudness ($\sim$ 0 dB). The magnitude spectrogram $X$ is computed by the STFT using a 64 ms sliding Blackman window with hop

[1] https://sites.google.com/site/unvoicedsoundseparation/mir-1k

size 30 ms and zero-padded to produce a feature vector of length 900 at each time step. The source separation quality is evaluated with the BSS Eval toolbox[2] using the standard metrics SDR, SIR and SAR that measure for each channel the ratios of source to distortion, interference and artifacts respectively [31]. For each model and singer combination, we use a random search on predefined intervals to select the hyperparameters that maximize the mean SDR on the validation set; final performance is reported on the test set.

## 6. RESULTS

To illustrate the effectiveness of our temporally constrained model, we first perform source separation experiments on a synthetic dataset of two sawtooth wave sources of different amplitudes and randomly shifted along both dimensions. Figure 2 shows an example of such sources (Fig. 2(a–b)), along with the sources estimated by supervised NMF with either no temporal constraint (Fig. 2(c–d)) or with an RNN with the cosine distance cost (Fig. 2(e–f)). While this problem is provably unsolvable for NMF alone or with simple temporal smoothing (eq. 17), the RNN-constrained model successfully separates the two mixed sources. This extreme example demonstrates that temporal constraints become crucial when the *content* of each time frame is not sufficient to distinguish each source.



(a) Source 1 (b) Source 2

(c) Estimated 1, NMF (d) Estimated 2, NMF

(e) Estimated 1, RNN (f) Estimated 2, RNN

**Fig. 2**. Toy example: separation of sawtooth wave sources of different amplitudes (a–b) using supervised NMF with either no prior (c–d) or an RNN with the cosine distance cost (e–f).

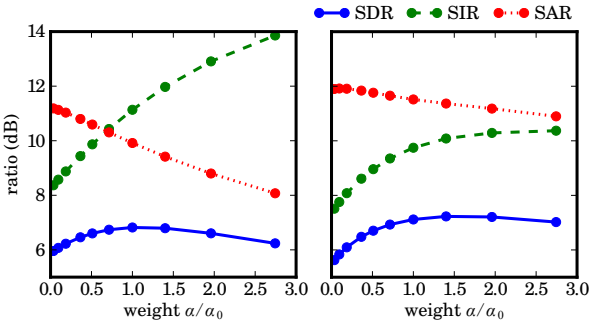
Source separation results on the MIR-1K dataset are presented in Table 1 for supervised (top) and semi-supervised[3] (bottom) NMF ($K = 15$). The RNN-based models clearly outperform the baselines in SDR and SIR for both sources with a moderate degradation in SAR. To illustrate the trade-off between the suppression of the unwanted source and the reduction of artifacts, we plot in Figure 3 the performance metrics as a function of the weight $\alpha/\alpha_0$ of the RNN-RBM model, where $\alpha_0 \in [10, 20]$ is the hyperparameter value selected on the validation set. This inherent trade-off was also observed elsewhere [15]. Overall, the observed improvement in SDR is indicative of a better separation quality.

[2] http://bass-db.gforge.inria.fr/bss_eval/

[3] Only the singing voice channel is supervised in this case.

| Model | SDR | | SIR | | SAR | |
|---|---|---|---|---|---|---|
| | acc. | sing. | acc. | sing. | acc. | sing. |
| NMF | 5.04 | 5.05 | 7.75 | 7.59 | 10.00 | 10.25 |
| NMF-sm | 6.08 | 5.59 | 8.77 | 7.42 | **10.96** | **11.93** |
| RNN | 6.13 | 5.80 | 9.46 | 7.79 | 10.34 | 11.52 |
| RNN-RBM | **6.83** | **7.12** | **11.25** | **9.75** | 9.86 | 11.52 |
| NMF | 5.20 | 3.58 | 9.54 | 4.95 | 8.80 | 11.43 |
| NMF-sm | 5.57 | 3.71 | 9.48 | 4.94 | **9.57** | 11.84 |
| RNN | 5.94 | 3.70 | 10.49 | 4.86 | 9.36 | **12.07** |
| RNN-RBM | **6.16** | **5.05** | **11.81** | **7.12** | 9.04 | 10.59 |

**Table 1**. Audio source separation performance on the MIR-1K test set obtained via singer-dependent supervised (top) and semi-supervised (bottom) NMF with either no prior, simple temporal smoothing, an RNN (eq. 11) or the RNN-RBM.




(a) Accompaniment (b) Singing voice

**Fig. 3**. Source separation performance trade-off on the MIR-1K test set by supervised NMF with an RNN-RBM model weighted by $\alpha$, where $\alpha_0$ maximizes the validation SDR.

## 7. CONCLUSION

We have presented a framework to leverage high-level information during feature extraction by incorporating an RNN-based prior inside the NMF decomposition. While the combined approach surpasses the baselines in realistic audio source separation settings, it could be further improved by employing a deep bidirectional RNN with multiplicative gates [19], replacing the Gaussian RBMs with the recently developed tractable distribution estimator for real-valued vectors RNADE [32, 17], implementing an EM-like algorithm to jointly train the NMF and RNN models, and transitioning to a universal speech model for singer-independent source separation [33].

## 8. REFERENCES


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

[2] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *ISMIR*, 2006.

[3] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 18, no. 3, pp. 528–537, 2010.

[4] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Discriminative non-negative matrix factorization for multiple pitch estimation.," in *ISMIR*, 2012, pp. 205–210.

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

[6] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation," in *ICASSP*, 2008.

[7] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *INTERSPEECH*, 2008.

[8] C. Feṿotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *ICASSP*, 2011.

[9] J. Nam, G. J. Mysore, and P. Smaragdis, "Sound recognition in mixtures," in *LVA/ICA*, 2012.

[10] C. Feṿotte, J. Le Roux, and J. R. Hershey, "Non-negative dynamical system with application to speech and audio," in *ICASSP*, 2013.

[11] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *ICASSP*, 2013.

[12] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *WASPAA*, 2009.

[13] M. Nakano, J. Le Roux, H. Kameoka, Y. Kitano, N. Ono, and S. Sagayama, "Nonnegative matrix factorization with Markov-chained bases for modeling time-varying patterns in music spectrograms," in *LVA/ICA*, 2010.

[14] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. on Acoustics, Speech, and Lang. Proc.*, vol. 21, no. 5, pp. 998–1011, 2013.

[15] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *LVA/ICA*, 2010, pp. 140–148.

[16] D. E. Rumelhart, G. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Dist. Proc.*, pp. 318–362. MIT Press, 1986.

[17] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *ICML 29*, 2012.

[18] T. Mikolov, A. Deoras, S. Kombrink, L. Burget, and J. Cernocký, "Empirical evaluation and combination of advanced language modeling techniques.," in *INTERSPEECH*, 2011, pp. 605–608.

[19] A. Graves, A.-R Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[20] Y. Bengio, N. Boulanger-Lewandowski, and R. Pascanu, "Advances in optimizing recurrent networks," in *ICASSP*, 2013.

[21] I. Sutskever, G. Hinton, and G. Taylor, "The recurrent temporal restricted Boltzmann machine," in *NIPS 20*, 2008, pp. 1601–1608.

[22] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS 13*, 2001.

[23] P. O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing*, 2002, pp. 557–565.

[24] C. J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[25] M. Welling, M. Rosen-Zvi, and G. Hinton, "Exponential family harmoniums with an application to information retrieval," in *NIPS 17*, 2005, pp. 1481–1488.

[26] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[27] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

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

[29] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "High-dimensional sequence transduction," in *ICASSP*, 2013.

[30] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Audio chord recognition with recurrent neural networks," in *ISMIR*, 2013.

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

[32] B. Uría, I. Murray, and H. Larochelle, "RNADE: The real-valued neural autoregressive density-estimator," in *NIPS 26*, 2013.

[33] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *ICASSP*, 2013.