NON-NEGATIVE DYNAMICAL SYSTEM WITH APPLICATION TO SPEECH AND AUDIO

Cédric Févotte[*][1], *Jonathan Le Roux*[2], *John R. Hershey*[2]

[1]Laboratoire Lagrange (CNRS, OCA & University of Nice), Parc Valrose, 06000 Nice, France
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA

ABSTRACT

Non-negative data arise in a variety of important signal processing domains, such as power spectra of signals, pixels in images, and count data. This paper introduces a novel non-negative dynamical system (NDS) for sequences of such data, and describes its application to modeling speech and audio power spectra. The NDS model can be interpreted both as an adaptation of linear dynamical systems (LDS) to non-negative data, and as an extension of non-negative matrix factorization (NMF) to support Markovian dynamics. Learning and inference algorithms were derived and experiments on speech enhancement were conducted by training sparse non-negative dynamical systems on speech data and adapting a noise model to the unknown noise condition. Results show that the model can capture the dynamics of speech in a useful way.

Index Terms— non-negative dynamical system (NDS), linear dynamical system (LDS), multiplicative innovations, non-negative matrix factorization (NMF), source separation.

1. INTRODUCTION

Non-negative sequences arise in a variety of important signal processing domains, such as power spectra of signals, pixels in images, and sequences of count data. This paper introduces a novel dynamical system for non-negative data, and describes its application to speech and audio modeling.

This work bridges two active fields, dynamical systems and non-negative matrix factorization (NMF). Dynamical systems are a long-standing area of research with applications in many scientific fields. A large body of literature is devoted to the case of linear dynamical systems (LDS), which describe an observed sequence $v_n \in \mathbb{R}^F$, indexed by $n \in [1, ..., N]$, via latent variables $h_n \in \mathbb{R}^K$, according to the equations:

$$h_n = Ah_{n-1} + \xi_n, \tag{1}$$

$$v_n = Wh_n + \epsilon_n, \tag{2}$$

where A and W are matrices of dimensions $K \times K$ and $F \times K$, respectively, and ξ_n and ϵ_n are independent, additive, and typically Gaussian random variables. Eq. (1) describes the dynamics of the state variable h_n and Eq. (2) describes the observation model. A, W, ξ_n and ϵ_n are referred to as state-transition matrix, dictionary, state innovation and data innovation, respectively. The LDS model does not naturally apply to the case where h_n and v_n are *non-negative*.

Besides, NMF is a more recent research area that has attracted a lot of attention in signal processing and machine learning communities since the publication of the seminal paper [1]. In the general case, NMF is the problem of finding an approximation of the form $V \approx WH$ where V and H are non-negative matrices. This approximation is generally obtained by minimizing a cost function $D(V|WH)$ that measures the dissimilarity between V and WH. In some settings, the columns of V form a sequence $v_1 \dots v_N$ with evolving dynamics, in the sense of statistical dependencies between elements in the sequence, that standard forms of NMF will fail to capture. Our work brings probabilistic dynamics to NMF, comparable to that of the traditional LDS.

The discrete-state Hidden Markov Model (HMM) is another dynamical system that has been commonly used to handle dynamics of speech, most famously in automatic speech recognition [2], but also in speech synthesis [3] as well as in speech separation [4]. In this setting, the speech features are usually taken to be cepstral coefficients or other log-spectrum-based features. However, HMMs lead to combinatorial complexity due to the discrete state-space, especially in the co-occurrence of several speakers. Because of the discreteness of the state space and the state-conditional independence of adjacent frames, HMMs also famously do not easily handle gain adaptation and continuity over time. In contrast, standard NMF solves both the computational cost (of linear complexity per iteration) and gain adaptation problems (through H), but it does not handle continuous dynamics. By bringing continuous dynamics to an NMF-like formulation, we hope to obtain the best of both worlds.

The proposed non-negative dynamical system (NDS) is given by

$$h_n = Ah_{n-1} \circ \xi_n, \tag{3}$$

$$v_n = Wh_n \circ \epsilon_n, \tag{4}$$

where all variables are non-negative with the same dimensions as above, and '$\circ$' denotes element-wise multiplication. The multiplicative innovations ξ_n and ϵ_n are non-negative random variables. The observation model (4) operates similarly to standard NMF, whereas the latent dynamics (3) capture statistical dependency between frames similarly to LDS.

In Section 2, we complete the presentation of the proposed model (3)-(4) with additional statistical assumptions. In Section 3, we present a majorization-minimization (MM) algorithm for maximum a posteriori (MAP) estimation of W, H, A given V. Section 4 discusses how the proposed NDS applies to the modeling of speech spectra and reports speech enhancement results.

2. MODEL

2.1. Statistical assumptions

We assume the state and data innovations to be independent and identically distributed (i.i.d), spatially independent and Gamma distributed, such that $p(\xi_1, \dots, \xi_N) = \prod_{k,n} G(\xi_{kn}|\alpha_k, \beta_k)$ and $p(\epsilon_1, \dots, \epsilon_N) = \prod_{f,n} G(\epsilon_{fn}|\nu_f, \delta_f)$, where α_k, β_k, ν_f and δ_f are positive scalars and G refers to the Gamma distribution, with probability density function (pdf) $G(x|\alpha, \beta) = \beta^\alpha/\Gamma(\alpha)\, x^{\alpha-1}\mathrm{e}^{-\beta x}$. It

[*]The research presented in this paper was conducted while C. Févotte was a visiting scientist at MERL over Summer 2012.

follows that v_n and h_n are conditionally Gamma distributed, such that

$$p(h_n|Ah_{n-1}) = \prod_k G(h_{kn}|\alpha_k, \beta_k / \sum_j a_{kj} h_{j(n-1)}), \quad (5)$$

$$p(v_n|Wh_n) = \prod_f G(v_{fn}|\nu_f, \delta_f / \sum_k w_{fk} h_{kn}). \quad (6)$$

The former holds for $n > 1$ only and we assume independent scale-invariant Jeffreys prior for $n = 1$, i.e., $p(h_{k1}) \propto 1/h_{k1}$. The expectation of the state variables and data under the model are

$$\mathbb{E}\left(h_{kn}|Ah_{n-1}\right) = \frac{\alpha_k}{\beta_k} \sum_j a_{kj} h_{j(n-1)}, \quad (7)$$

$$\mathbb{E}\left(v_{fn}|Wh_n\right) = \frac{\nu_f}{\delta_f} \sum_k w_{fk} h_{kn}. \quad (8)$$

For simplicity, we will assume $\nu_f = \delta_f = \delta$, so that $\mathbb{E}(V) = WH$, which is a natural assumption underlying most NMF settings (see for example [5]). Under this assumption, the negative log-likelihood $-\log p(V|WH)$ is essentially the Itakura-Saito (IS) divergence from V to WH, as explained in [6]. The expression of $p(h_n|Ah_{n-1})$ reveals a scale ambiguity between β_k and $\{a_{kj}\}_j$, which we fix by setting $\beta_k = \alpha_k$, implying that $\mathbb{E}(h_n|Ah_{n-1}) = Ah_{n-1}$.

The multiplicative and non-negative state and data innovations preserve the non-negativity in the generative model. Additive real-valued Gaussian innovations, as used in LDS, would fail to do so. Additive non-negative innovations would ensure non-negativity, but would only allow monotonically increasing values. In contrast, for multiplicative innovations that can take values both lower and greater than one, the coefficients of v_n and h_n are allowed to increase and decrease.

2.2. Related work

To the best of our knowledge there is no prior work about the NDS (3)-(4) proposed in this paper. There is some literature on non-negative dynamical systems, where the state-space matrices, state and observation vectors are assumed non-negative. However, to the best of our knowledge, only non-negative additive perturbations have been considered, see, e.g., [7, 8]. Furthermore most of the literature we found on the topic is devoted to the theoretical properties of these systems (stability, observability) rather than inference in or application of these models. The closest to our model is perhaps [9], which studies the conditions for existence, uniqueness and stability of the system defined by Eq. (3) alone, but does not provide algorithms nor procedures for inference.

A special case of the NDS (3)-(4) has been addressed in the NMF literature. In [6, 10], algorithms were derived for the special case $A = I_K$, where I_K denotes the identity matrix of dimension K, and $\delta = 1$ (multiplicative exponentially-distributed observation innovation). In these papers, the state innovation was arbitrarily assumed either Gamma or inverse-Gamma with mode obtained at 1, i.e., the mode of $p(h_{kn}|h_{k(n-1)})$ is obtained at $h_{kn} = h_{k(n-1)}$. The resulting method, coined smooth Itakura-Saito NMF, allows to regularize the individual rows of H temporally, assuming mutual independence of the rows. The proposed model goes an important step forward by lifting the mutual independence assumption, which is generally not a realistic one. Indeed, the activation of a pattern at frame n is also likely to correlate (or anti-correlate) with the activation of other patterns at frame $n-1$. This is what the proposed model achieves through the introduction of the state-transition matrix A in Eq. (3).

A preliminary attempt at introducing dynamics into NMF has been made by Smaragdis *et al.*, for speech denoising [11] and sound classification [12]. They compute standard NMF decompositions from spectral training data describing each sound class (e.g., in the denoising setting, speech and noise) and compute the average temporal dynamics of the returned matrices H for each class *a posteriori*. Test sounds are then decomposed onto the learned spectral patterns, using a regularization step that employs the precomputed temporal statistics of each class. In particular [12] employs an ad-hoc forward-backward smoothing of the activations. The work presented in this paper pursues a more formal approach, and proposes a well-posed statistical model for non-negative data along with principled algorithms for inference.

3. MAXIMUM A POSTERIORI ESTIMATION

We have derived a majorization-minimization (MM) algorithm for MAP estimation of the parameters W, A, H given user-defined values of the remaining parameters δ and $\{\alpha_k\}_k$. In the following, we only present the main steps of the procedure to meet the space limitations constraints. The MAP objective function is defined by $C(W, H, A) = -\log p(V|WH) - \log p(H|A)$. Our MM algorithm is a block-coordinate descent algorithm that updates W, H and A individually and conditionally upon the current values of the other parameters. The algorithm alternates between forming an upper bound of the objective function at the current parameter settings, and optimizing parameters to minimize this bound.

Let $\tilde{W}$, $\tilde{H}$ and $\tilde{A}$ denote the parameter values at the current iteration and consider for example the update of W. Denote for example $F(W) = C(W, \tilde{H}, \tilde{A})$ the function to be minimized w.r.t W. The first step of the MM algorithm consists in building an upper bound $G(W, \tilde{W})$ of $F(W)$ which is tight for $W = \tilde{W}$, i.e., $F(W) \leq G(W, \tilde{W})$ for all W and $G(\tilde{W}, \tilde{W}) = F(\tilde{W})$. The second step consists in minimizing the bound w.r.t W, producing a valid descent algorithm. Indeed, at iteration $i + 1$, it holds by construction that $F(W^{(i+1)}) \leq G(W^{(i+1)}, W^{(i)}) \leq G(W^{(i)}, W^{(i)}) = F(W^{(i)})$. The same principle applies to the updates of H and A.

The upper bounds can be derived using standard inequalities, namely Jensen's inequality for the convex parts of the objective functions and the tangent inequality for the concave parts. Using this strategy, the updates of W and A are multiplicative and given by

$$w_{fk} = \tilde{w}_{fk} \sqrt{\frac{\sum_{n=1}^{N} h_{kn} v_{fn} / \tilde{v}_{fn}^2}{\sum_{n=1}^{N} h_{kn} / \tilde{v}_{fn} + \lambda}}, \quad (9)$$

$$a_{kj} = \tilde{a}_{kj} \sqrt{\frac{\beta_j \sum_{n=2}^{N} h_{j(n-1)} h_{kn} / \tilde{g}_{kn}^2}{\alpha_j \sum_{n=2}^{N} h_{j(n-1)} / \tilde{g}_{kn}}}, \quad (10)$$

where $\tilde{v}_{fn} = \sum_k \tilde{w}_{fk} h_{kn}$, $\tilde{g}_{kn} = \sum_j \tilde{a}_{kj} h_{j(n-1)}$ and λ is a constant that prevents degenerate solutions such that $\|W\| \to \infty$ and $\|H\| \to 0$. Owing to the Markovian structure of H, adjacent columns are coupled in the optimization. We have employed a left-to-right block-coordinate descent approach that updates h_n at iteration i conditionally on $h_{n-1}^{(i)}$ and $h_{n+1}^{(i-1)}$, for $1 < n < N$. With this approach, updates of h_{kn} are available in closed form and merely involve rooting a polynomial of order 2, such that

$$h_{kn} = \frac{\sqrt{q_{kn}^2 - 4 p_{kn} r_{kn}} - q_{kn}}{2 p_{kn}}, \quad (11)$$

where $p_{kn} = \delta \sum_f \frac{w_{fk}}{\tilde{v}_{fn}} + \sum_j \alpha_j \frac{a_{jk}}{\tilde{g}_{j(n+1)}} + \frac{\beta_k}{g_{kn}}$, $q_{kn} = 1 - \alpha_k$, $r_{kn} = -\tilde{h}_{kn}^2 \Big(\delta \sum_f w_{fk} \frac{v_{fn}}{\tilde{v}_{fn}^2} + \sum_j \beta_j \frac{a_{jk} h_{j(n+1)}}{\tilde{g}_{j(n+1)}^2}\Big)$, and $\tilde{v}_{fn} = \sum_k w_{fk} \tilde{h}_{kn}$, $\tilde{g}_{k(n+1)} = \sum_j a_{kj} \tilde{h}_{jn}$, $g_{kn} = \sum_j a_{kj} h_{j(n-1)}$.

4. APPLICATION TO SPEECH

4.1. Spectral modeling of speech with NDS

When $\delta = 1$, such that ξ_{kn} is exponentially distributed, Eq. (4) can be related to a generative model of the power spectrogram in the following Gaussian composite model (GCM). Let x_{fn} denote the complex-valued short-time Fourier transform (STFT) of some time domain audio signal, where f is a frequency bin index and n indexes time frames. The GCM is defined by $x_{fn} = \sum_k c_{fkn}$ and $c_{fkn} \sim N_c(0, w_{fk} h_{kn})$, where $N_c(0, \lambda)$ refers to the circular complex Gaussian distribution with zero mean. The latent components $\{c_{fkn}\}$ can trivially be marginalized from the generative model, yielding $x_{fn} \sim N_c(0, \sum_k w_{fk} h_{kn})$. It follows that the power spectrogram $v_{fn} = |x_{fn}|^2$ of x_{fn} is exponentially distributed with mean $\sum_k w_{fk} h_{kn}$, and can thus be written as Eq. (4) with $\delta = 1$. Note that, when necessary, minimum mean squares estimate (MMSE) of the components can be obtained by Wiener filtering and given by

$$\hat{c}_{fkn} = \frac{w_{fk} h_{kn}}{\sum_j w_{fj} h_{jn}} x_{fn}. \tag{12}$$

The GCM has found successful applications in audio source separation and music transcription [6, 13, 14, 15, 16], and generalizes earlier two-component models used in spectral-based audio denoising, e.g., [17]. In contrast with Gaussian mixture models (GMMs) or HMMs, which are prevalent in speech log-spectral modeling and where each data frame is assumed to be in one among many possible states each characterized by a given covariance, the GCM assumes that each data frame is a sum of zero-mean Gaussian-distributed components. In this paper the GCM is used as an observation process for the NDS model.

4.2. Speech enhancement with NDS

We consider a speech enhancement scenario where the time-domain data x_t is a clean speech signal s_t corrupted by additive noise b_t, such that

$$x_t = s_t + b_t, \tag{13}$$

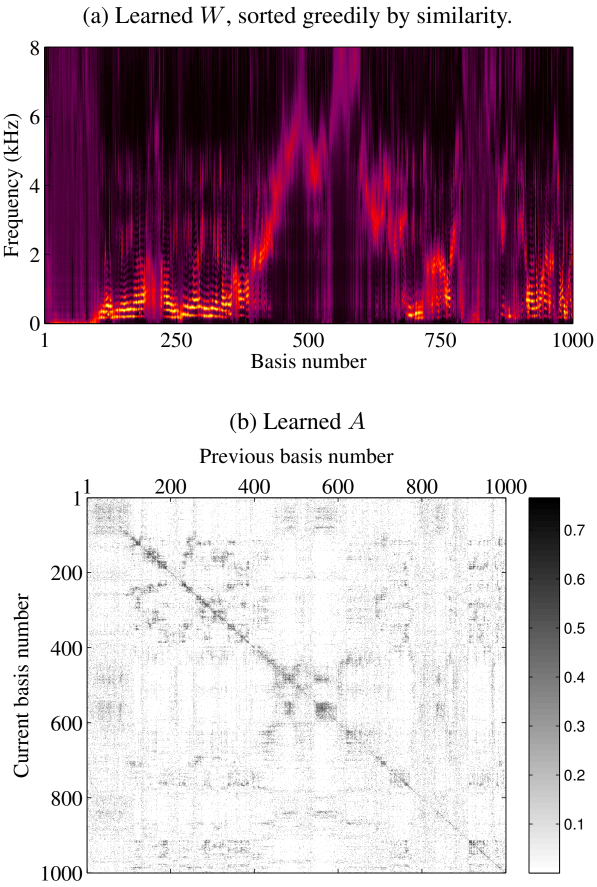
and we wish to produce a speech estimate $\hat{s}_t$ of s_t. Given a trained NDS model of speech (W^{train}, A^{train}), representative of the unseen source s, and given the corrupted data x, we estimate the nonnegative decomposition

$$V \approx W^{\text{train}} H + W^{\text{noise}} H^{\text{noise}} \tag{14}$$

of the power spectrogram V of the noisy observation x, where $W^{\text{train}} H$ represents the speech spectrogram estimate and $W^{\text{noise}} H^{\text{noise}}$ the noise spectrogram estimate. We then reconstruct the time-domain source estimate $\hat{s}$ by MMSE estimation, which amounts to Wiener filtering in our model. That is, we take the inverse STFT of

$$\hat{S} = \frac{W^{\text{train}} H}{W^{\text{train}} H + W^{\text{noise}} H^{\text{noise}}} \circ X, \tag{15}$$

where the fraction bar is here elementwise.


(a) Learned W, sorted greedily by similarity.

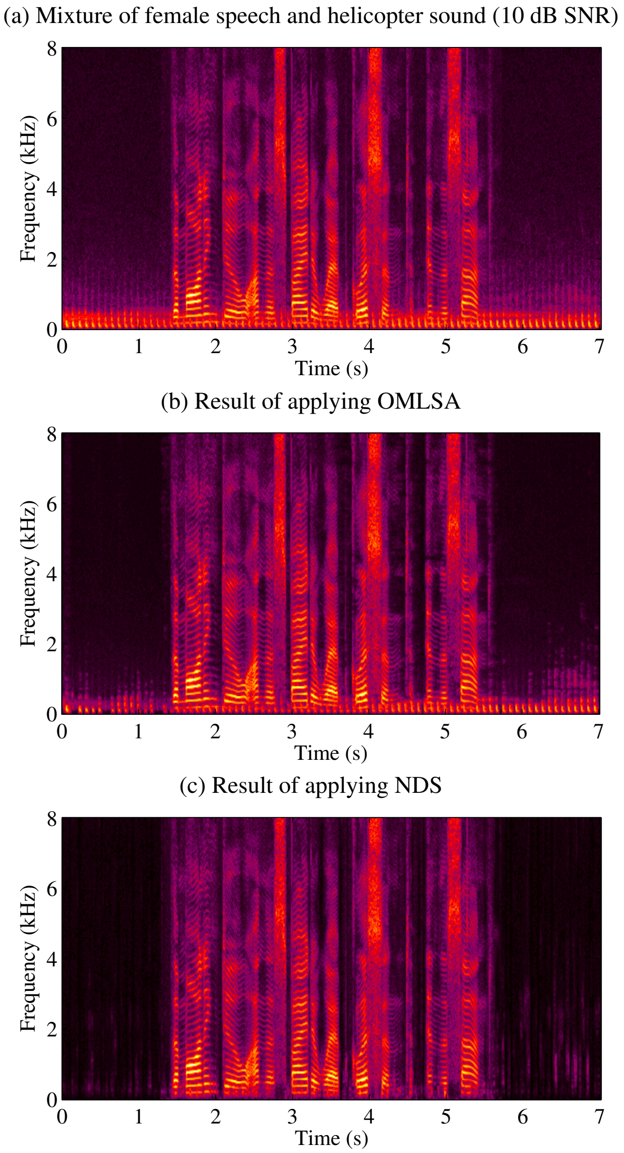
(b) Learned A

Fig. 1. Training NDS on female speech (frame length: 512).

In our approach, we assume the source and noise STFTs to follow a GCM, such that $s_{fn} \sim N_c(0, \sum_k w_{fk}^{\text{train}} h_{kn})$ and $b_{fn} \sim N_c(0, \sum_k w_{fk}^{\text{noise}} h_{kn}^{\text{noise}})$. H follows the dynamics of Eq. (3) with transition matrix A^{train}. We recall that W^{train}, A^{train} are fixed variables, learned in a training phase. The matrices W^{noise} and H^{noise} are here assumed unknown and with no particular structure, and learned from the data. With these assumptions, the decomposition (14) shall be obtained by minimizing the divergence $D_{IS}(V|W^{\text{train}}H + W^{\text{noise}}H^{\text{noise}})$ penalized by the dynamical term $-\log p(H|A^{\text{train}})$, w.r.t H, W^{noise} and H^{noise}. This can be achieved with a MM algorithm using minor modifications of the derivations presented in Section 3. Besides the use of the IS divergence and the dynamical penalty term, the proposed procedure resembles the semi-supervised NMF setting described in [18].

4.3. Enhancement results

The sampling rate was 16 kHz. Time-frequency analysis was performed using frame lengths of 512, 640, 800 and 960 samples, using for each length a 50% overlap and a sine window for analysis and re-synthesis. For each window length, NDS models were trained separately for male and female speech, each on 1000 utterances (about 50 minutes) from the TIMIT training set. The number of bases was set to $K = 1000$ and the Gamma distribution parameter to $\alpha_k = \alpha = 0.01$. We show in Fig. 1 examples of W and A trained on female speech with a frame length of 512. The low setting



(a) Mixture of female speech and helicopter sound (10 dB SNR)

(b) Result of applying OMLSA

(c) Result of applying NDS

Fig. 2. Speech enhancement results.

of α strongly encourages sparsity in H, thus encouraging a holistic representation of the spectrum, and a sparse transition matrix. However, sparsity can lead to jitter in the estimated signals, a potential trade-off that will need to be investigated in future work.

We evaluated the enhancement algorithm described above on mixtures of 10 speech files by different speakers (5 male and 5 female) from the TIMIT test set with 15 environmental texture sounds from [19] at 3 different input signal-to-noise ratios (SNR), for a total of 450 mixtures. The training and test sets had disjoint sets of speakers. The texture sounds include a wide variety of environmental sounds such as fire, bees, water stream, helicopter, applause, babble noise, shaking paper, etc. The speech test files were 1.8 s to 4.5 s long, and were added in the middle of the 7 s long environmental sounds. The number of noise bases was set to $K^{\text{noise}} = 2$. For each mixture, we assumed the gender known and used the NDS speech model for that gender.

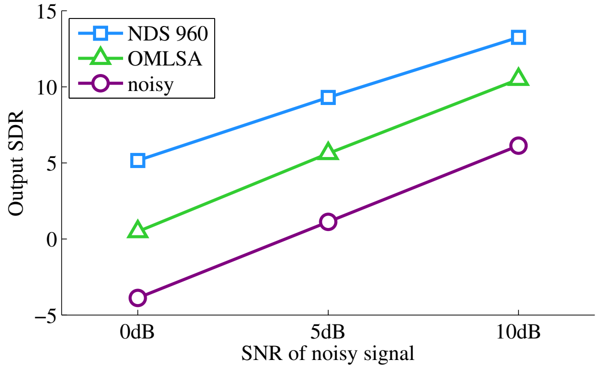


Fig. 3. Signal to distortion ratios for reconstructed speech with NDS (frame length 960, corresponding to 60 ms) and with OMLSA, and for the noisy speech (used as a baseline).

For comparison, we also show results for the state-of-the-art algorithm combining Optimally-Modified Log Spectral Amplitude Estimator and Improved Minima Controlled Recursive Averaging [20, 21], denoted as OMLSA. To illustrate the behavior of the algorithms, we show in Fig. 2 the spectrogram of a mixture of speech by a female speaker and a helicopter sound at 10 dB SNR, and the output of OMLSA and NDS on that mixture. NDS is able to suppress the non-stationary helicopter noise, while OMLSA fails to do so. The outputs of the algorithms were also quantitatively evaluated using the `bss_eval` toolbox [22], treating our denoising problem as a source separation problem where s is the target source and b is the interfering source, and the perception of speech quality (PESQ) measure [23].

The `bss_eval` results are given in terms of signal-to-distortion ratio (SDR) and shown in Fig. 3. The proposed NDS algorithm with 60 ms windows significantly outperforms OMLSA for all input SNRs. The results for other frame lengths were similar in terms of SDR. When measured by PESQ, in contrast, the scores generally increased with window size, but yielded no significant improvement compared to OMLSA for the window sizes investigated. In informal listening tests, the NDS model was very good at removing a wide range of non-stationary noises, but suffered from a tendency to leave behind residual speech-like sounds. In particular, due to the fact that training data contained lip smacks, breath sounds, etc, noises ressembling these were often passed unsuppressed into the speech estimate. We are currently investigating whether these artifacts can be controlled by alternative parameterizations of the model.

5. CONCLUSIONS

We presented a novel non-negative dynamical system called NDS to model sequences of non-negative data, explained its relationships with previous work, derived an efficient MAP estimation algorithm, and explained how it can be applied to speech and audio modeling. In preliminary experiments on a speech enhancement task with real environmental sounds, the proposed NDS algorithm reached similar performance in terms of PESQ scores compared to the state-of-the-art algorithm, and significantly outperformed the state of the art in terms of SDR. Future works include testing more thoroughly the potential of the proposed model on non-negative data of different modalities, developing extensions enabling the sharing of bases across states, and investigating usage of NDS as a noise model.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects with nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[2] B. H. Juang and L. R. Rabiner, "Hidden Markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, vol. 3, pp. 1315–1318.

[4] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[5] V. Y. F. Tan and C. Févotte, "Automatic relevance determination in nonnegative matrix factorization with the beta-divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press.

[6] 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.

[7] J. M. van den Hof, "Realization of positive linear systems," *Linear Algebra and its Applications*, vol. 256, no. 0, pp. 287 – 308, 1997.

[8] W. M. Haddad and V. Chellaboina, "Stability and dissipativity theory for nonnegative dynamical systems: a unified analysis framework for biological and physiological systems," *Nonlinear Analysis: Real World Applications*, vol. 6, no. 1, pp. 35–65, 2005.

[9] J. Stachurski, "Economic dynamical systems with multiplicative noise," *Journal of Mathematical Economics*, vol. 39, no. 1-2, pp. 135–152, 2003.

[10] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[11] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2008, pp. 4029–4032.

[12] J. Nam, G. J. Mysore, and P. Smaragdis, "Sound recognition in mixtures," in *Proc. International Conference on Latent Variable Analysis & Independent Component Analysis (LVA/ICA)*, Tel-Aviv, Israel, 2012.

[13] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by nonnegative sparse coding of power spectra," in *Proc. 5th International Symposium Music Information Retrieval (ISMIR)*, Barcelona, Spain, Oct. 2004, pp. 318–325.

[14] L. Benaroya, R. Gribonval, and F. Bimbot, "Non negative sparse representation for Wiener based source separation with a single sensor," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 2003, pp. 613–616.

[15] R. M. Parry and I. Essa, "Phase-aware non-negative spectrogram factorization," in *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, London, UK, Sep. 2007, pp. 536–543.

[16] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 550–563, Mar. 2010.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. 7th International Conference on Independent Component Analysis and Signal Separation (ICA)*, London, UK, Sep. 2007.

[19] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, 2011.

[20] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, 2002.

[21] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

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

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 749–752.