NON-NEGATIVE SOURCE-FILTER DYNAMICAL SYSTEM FOR SPEECH ENHANCEMENT

Umut Şimşekli,[1] *Jonathan Le Roux,*[2] *John R. Hershey*[2]

[1]Boğaziçi University, Dept. of Computer Engineering, 34342, Bebek, İstanbul, Turkey
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
umut.simsekli@boun.edu.tr, {leroux, hershey}@merl.com

ABSTRACT

Model-based speech enhancement methods, which rely on separately modeling the speech and the noise, have been shown to be powerful in many different problem settings. When the structure of the noise can be arbitrary, which is often the case in practice, model-based methods have to focus on developing good speech models, whose quality will be key to their performance. In this study, we propose a novel probabilistic model for speech enhancement which precisely models the speech by taking into account the underlying speech production process as well as its dynamics. The proposed model follows a source-filter approach where the excitation and filter parts are modeled as non-negative dynamical systems. We present convergence-guaranteed update rules for each latent factor. In order to assess performance, we evaluate our model on a challenging speech enhancement task where the speech is observed under non-stationary noises recorded in a car. We show that our model outperforms state-of-the-art methods in terms of objective measures.

***Index Terms*—** source-filter model, non-negative dynamical system, non-negative matrix factorization, speech enhancement, source separation

1. INTRODUCTION

Speech enhancement methods attempt to improve the quality and intelligibility of speech that has been degraded by interfering noise or other processes. The aim is generally to recover the clean speech signal from a noisy mixture, where the mixture is assumed to be the sum of the speech signal and a noise signal.

Model-based speech enhancement methods aim to express the speech and the noise spectra using statistical models. For situations where the noise is stationary or slowly varying, relatively simple models of both speech and noise can be very effective [1,2]. In more general settings, where the structure of the noise is unpredictable, the quality of the speech model plays a key role in speech enhancement performance. In this case, a *semi-supervised* approach can be taken, where the speech model is estimated on speech training data and the noise model is estimated during the enhancement process.

Model-based speech enhancement methods differ in terms of the basic modeling distributions and strategy, the feature domain used for modeling, and the extent to which structure such as temporal dynamics and speech production properties are modeled.

In terms of modeling strategy, two broad approaches exist: one based on discrete state modeling such as Gaussian mixture models (GMMs) and hidden Markov models (HMMs) versus methods using continuously-weighted combinations of basis functions, such as non-negative matrix factorizations (NMF) [3] and their extensions. The general trade-off is that discrete-state approaches [4, 5] can be more precise, especially in their temporal dynamics, whereas continuous approaches [6, 7] can be more flexible with respect to gain and subspace variability.

Feature domains such as the complex spectrum, power spectrum, and log power spectrum have been used for speech enhancement. Each domain introduces a trade-off between the ease of modeling the signals, and that of modeling the interaction between signals that are mixed together [8]. In feature domains where the interaction between speech and noise is additive, isolating the phonetic content of the speech signal can be difficult. This is because phonetic content is imparted to speech by the filtering effect of the vocal tract, which is approximately multiplicative in the power spectrum. In the log spectrum domain the vocal tract filter is additive, but the effect of noise is nonlinear, and compensating for it becomes difficult.

Many systems based on single-frame modeling of the speech spectrum have been investigated, including log spectrum GMMs [9], or other spectral mixture models [10], as well as power-spectrum domain NMF models. Such models tend to be susceptible to transients and in general could benefit from the known dynamical structure present in speech signals: the evolution of phonetic and pitch processes are governed by linguistic constraints as well as constraints on speech production. Models have been proposed that incorporate such structure, such as temporal dynamics and source-filter modeling. Discrete state models, such as HMMs, represent dynamics using discrete state transitions over time [4, 11]. Continuous state Gaussian dynamical models, such as linear dynamical systems (LDSs), have long been studied [12], and recently rich models of continuous dynamics have been extended to the NMF family using gamma-distributed models [6,7] in models known as *non-negative dynamical systems* (NDSs). There have also been combinations with discrete dynamics and NMF observation models [13].

Knowledge of speech production mechanisms can also be exploited to impose powerful modeling constraints. Source-filter models represent the excitation source and the filtering of the vocal tract as separate factors [14]: the source corresponds to the excitation part of the signal which is mainly composed of vocal cord vibrations (voicing) having a particular pitch, turbulent air noise (fricatives), and air flow onset/offset sounds (stops), and their combinations. The filter corresponds to the influence of the vocal tract on the spectral envelope of the sound, as in the case of different vowels ('ah' versus 'ee') or differently modulated fricative modes ('s' versus 'sh'). Such a factorial strategy has been proposed in various domains [15–20]. In [5], factorial HMMs were used to model both the source and filter dynamics for speech separation, but otherwise there has been little work modeling dynamics of both factors.

We investigate a novel probabilistic model for speech enhancement that draws from many of the above approaches. The aim is

This research was conducted while Umut Şimşekli was an intern at MERL. The authors thank Dr. Cédric Févotte for fruitful discussions.

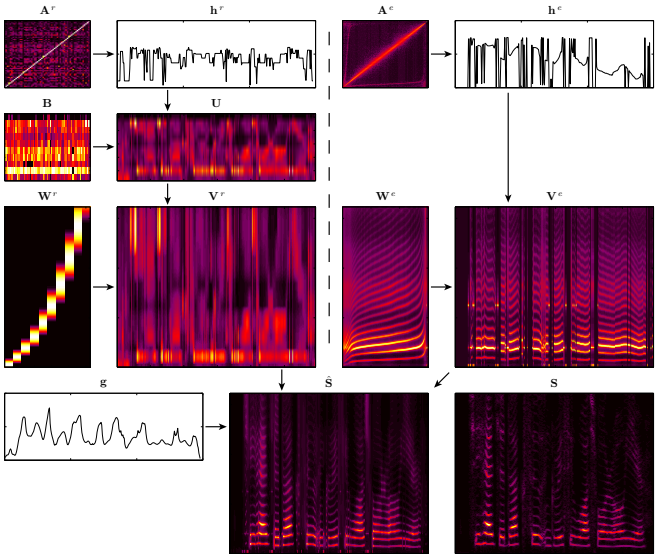

Fig. 1. Illustration of the proposed model. The power spectrum $\mathbf{S}$ is decomposed as a product of a filter part $\mathbf{V}^r$, an excitation part $\mathbf{V}^e$, and gains $\mathbf{g}$. The smooth overlapping filter dictionary $\mathbf{W}^r$ implicitly restricts $\mathbf{V}^r$ to capture the smooth envelope of the spectrum. $\mathbf{W}^e$ captures the spectral shapes of the excitation modes. $\hat{\mathbf{S}}$ is the model prediction: $\hat{s}_{fn} = g_n v^r_{fn} v^e_{fn}$.

to model the speech precisely by taking into account the underlying speech production process as well as its dynamics. The proposed model follows a source-filter approach where the excitation and filter parts are modeled as a dynamical system. The state is factorized into discrete components for the filter (i.e., phoneme) states and the excitation states, and a continuous state for the overall gain. Each of these is modeled as a Markov chain, leading to a hybrid between a factorial HMM and the non-negative dynamical system approach. Whereas the excitation states directly select excitation templates similarly to [20], the filter observation model follows that of *hierarchical NDS* (HNDS) model [7] to allow for richer variations.

We evaluate our model on a challenging speech enhancement task where the speech is observed under non-stationary car noises. We show that our model outperforms the state-of-the-art methods in terms of objective measures, and that the dynamics and the hierarchical filter model each contribute to better performance.

The rest of the paper makes use of the following notation: bold capital letters denote matrices (e.g., $\mathbf{A}$), $\mathbf{a}_j$ denotes the j^{th} column of $\mathbf{A}$, and a_{ij} denotes a single entry of $\mathbf{A}$. Similarly, bold small letters denote vectors (e.g., $\mathbf{a}$) and a_i denotes a single entry of $\mathbf{a}$.

2. THE MODEL

We propose a non-negative source-filter dynamical system (NSFDS) model. NSFDS models the complex spectrum $\mathbf{X} \in \mathbb{C}^{F \times N}$ as a conditionally zero-mean complex Gaussian distribution,

$$x_{fn} \sim \mathcal{N}_c(x_{fn}; 0, g_n v^r_{fn} v^e_{fn}), \tag{1}$$

whose variance is modeled as the product of a filter component v^r_{fn}, an excitation component v^e_{fn}, and a gain g_n, where f denotes the frequency index and n the frame index. The filter component aims to capture the time-varying structure of the phonemes, whereas the excitation component aims to capture time-varying pitch and other

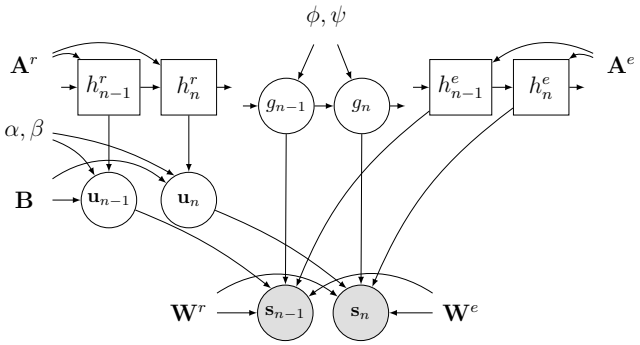

Fig. 2. Graphical representation of the proposed model. Circular nodes denote the continuous random variables, rectangular nodes denote the discrete random variables, and shaded nodes denote the observed variables. The arrows determine the conditional independence structure.

excitation modes of the speech. The gain component helps the model to track changes in amplitude.

This modeling approach is equivalent to assuming an exponential distribution over the power spectrum $s_{fn} = |x_{fn}|^2$, with $s_{fn} \sim \mathcal{E}(s_{fn}; 1/(g_n v^r_{fn} v^e_{fn}))$. Maximum likelihood estimation on this model is equivalent to minimizing the Itakura-Saito divergence between s_{fn} and $g_n v^r_{fn} v^e_{fn}$ [21].

For a given time frame n, the excitation component $\mathbf{v}^e_n$ is assumed to be a column of an excitation dictionary $\mathbf{W}^e \in \mathbb{R}_+^{F \times K_e}$:

$$v^e_{fn} = \prod_m w^{e\,[h^e_n = m]}_{fm}, \tag{2}$$

where $[\cdot]$ is the indicator function, i.e., $[x] = 1$ if x is true and 0 otherwise. Here, the discrete random variable $h^e_n \in \{1, \ldots, K_e\}$ is called 'excitation label' and determines the pitch and other excitation modes.

We model the filter component $\mathbf{V}^r$ as the multiplication of a predetermined filter dictionary $\mathbf{W}^r \in \mathbb{R}_+^{F \times K_r}$ and an activation matrix $\mathbf{U} \in \mathbb{R}_+^{K_r \times N}$, where we further restrict the domain of $\mathbf{U}$ in such a way that each column of $\mathbf{U}$ is a noisy realization of a column of an activation dictionary $\mathbf{B} \in \mathbb{R}_+^{K_r \times I_r}$:

$$\begin{aligned} v^r_{fn} &= \sum_k w^r_{fk} u_{kn}, \\ u_{kn} &= \Big(\prod_i b_{ki}^{[h^r_n = i]}\Big) \epsilon^u_{kn}, \qquad \epsilon^u_{kn} \sim \mathcal{G}(\epsilon^u_{kn}; \alpha, \beta). \end{aligned} \tag{3}$$

We call $h^r_n \in \{1, \ldots, I_r\}$ a 'phoneme label' and h^r_n determines the column of $\mathbf{B}$ that is chosen at time frame n. The gamma distribution $\mathcal{G}$ is defined using shape and inverse scale parameters.

In order to introduce continuous dynamics and enforce smoothness, we assume a gamma Markov chain on the gain variables $\mathbf{g}$:

$$g_n = \Big(g_{n-1}\Big)\epsilon^g_n, \qquad \epsilon^g_n \sim \mathcal{G}(\epsilon^g_n; \phi, \psi). \tag{4}$$

For simplicity, we constrain the innovations ϵ to have mean 1 by taking $\alpha = \beta$, $\phi = \psi$. Finally, we assume Markovian priors on the phoneme labels $\mathbf{h}^r$ and the excitation labels $\mathbf{h}^e$ in order to incorporate contextual information, with transition matrices $\mathbf{A}^r$ and $\mathbf{A}^e$:

$$\begin{aligned} h^r_n | h^r_{n-1} &\sim \prod_i \prod_j a^{r\,[h^r_n = i][h^r_{n-1} = j]}_{ij}, \\ h^e_n | h^e_{n-1} &\sim \prod_i \prod_j a^{e\,[h^e_n = i][h^e_{n-1} = j]}_{ij}. \end{aligned} \tag{5}$$

Table 1. Update rules for $\mathbf{U}$ and $\mathbf{g}$ for clean speech. Each variable can be updated at each iteration to $\frac{\sqrt{b^2-4ac}-b}{2a}$ with different a, b, and c values for each variable. Here, we define $\hat{s}_{fn} = g_n v^r_{fn} v^e_{fn}$.

	a	b	c
u_{kn}	$\sum_f \frac{w^r_{fk}}{v^r_{fn}} + \frac{\beta}{\prod_i b_{ki}^{[h^r_n=i]}}$	$1-\alpha$	$-u^2_{kn} \sum_f \frac{s_{fn}}{g_n v^e_{fn} \hat{s}^2_{fn}} w^r_{fk}$
$g_n\ (n=1)$	$(F+\phi)^2$	0	$-\left[\sum_f \frac{s_{fn}}{v^r_{fn} v^e_{fn}} + \psi g_{n+1}\right]^2$
$g_n\ (1<n<N)$	$\frac{\psi}{g_{n-1}}$	$F+1$	$-\left[\sum_f \frac{s_{fn}}{v^r_{fn} v^e_{fn}} + \psi \frac{g_n}{g_{n+1}}\right]$
$g_n\ (n=N)$	$\frac{\psi}{g_{n-1}}$	$F+1-\phi$	$-\left[\sum_f \frac{s_{fn}}{v^r_{fn} v^e_{fn}}\right]$

Note that the filter and excitation Markov chains could also be made interdependent to better model statistical relationships between the two, but here we leave them marginally independent. Making them dependent would increase the complexity of the model and the potential benefits remain to be explored.

Finally, we obtain the ultimate model by combining Eqs. 1-5. An illustration of the proposed NSFDS model is depicted in Fig. 1. The graphical models for the NSFDS model and related models are given in Fig. 2.

3. INFERENCE

In this section, we present convergence-guaranteed update rules for maximum a-posteriori (MAP) estimation in the proposed model. In particular, we use the majorization-minimization (MM) algorithm [22] which monotonically decreases the intractable MAP objective function by minimizing a tractable upper-bound constructed at each iteration. This algorithm is a block-coordinate descent algorithm which performs alternating updates of each latent factor given its current value and the other factors. For more details, the reader is referred to [22]. The MM algorithm yields the following updates for $\mathbf{B}$ and $\mathbf{W}^e$:

$$b_{ki} \leftarrow \frac{\beta \sum_n [h^r_n = i] u_{kn}}{\alpha \sum_n [h^r_n = i]}, \quad w^e_{fm} \leftarrow \frac{\sum_n [h^r_n = m] \frac{s_{fn}}{g_n v^r_{fn}}}{\sum_n [h^e_n = m]} \quad (6)$$

The updates of $\mathbf{U}$ and $\mathbf{g}$ involve finding roots of second order polynomials. The corresponding equations are given in Table 1. Finally, given all other variables, the optimal $\mathbf{h}^r$ and $\mathbf{h}^e$ can be computed via Viterbi algorithm at each iteration. The transition matrices $\mathbf{A}^r$ and $\mathbf{A}^e$ are estimated from the transition counts in the training data. A more detailed explanation of the update rules is provided in a supplementary document hosted on our project webpage [23].

4. SPEECH ENHANCEMENT EXPERIMENTS

4.1. Noisy Speech Model

We consider a mixture of speech with additive noise, which leads to a linear relationship in the complex spectrum domain, $x^{\text{mix}}_{fn} = x^{\text{speech}}_{fn} + x^{\text{noise}}_{fn}$. This avoids assuming additivity of the power spectra, an approximation made by many other methods. This is straightforward if the speech and the noise are both modeled with conditionally zero-mean complex Gaussian distributions:

$$x^{\text{speech}}_{fn} \sim \mathcal{N}_c(x^{\text{speech}}_{fn}; 0, v^{\text{speech}}_{fn}), \quad x^{\text{noise}}_{fn} \sim \mathcal{N}_c(x^{\text{noise}}_{fn}; 0, v^{\text{noise}}_{fn}). \quad (7)$$

Here, x^{speech}_{fn} is modeled by NSFDS, i.e., $v^{\text{speech}}_{fn} = g_n v^r_{fn} v^e_{fn}$ as defined in Eqs. 2-4. For the noise, we use smooth NMF (SNMF) [24], which is a simple and flexible model for non-stationary signals:

$$h^{\text{noise}}_{kn} = h^{\text{noise}}_{k(n-1)} \epsilon^h_{kn}, \qquad \epsilon^h_{kn} \sim \mathcal{G}(\epsilon^h_{kn}; \alpha^{\text{noise}}, \beta^{\text{noise}}),$$
$$v^{\text{noise}}_{fn} = \sum_k w^{\text{noise}}_{fk} h^{\text{noise}}_{kn}, \quad (8)$$

where v^{noise}_{fn} is assumed to be the product of a spectral dictionary $\mathbf{W}^{\text{noise}}$ and its corresponding activations $\mathbf{H}^{\text{noise}}$. SNMF is an extension of NMF that imposes a gamma Markov chain on the activations in order to enforce smoothness. Here, we set $\alpha^{\text{noise}} = \beta^{\text{noise}}$ to constrain the innovations ϵ^h_{kn} to have mean 1.

For each test case, we estimate the variables $\mathbf{h}^r$, $\mathbf{h}^e$, $\mathbf{U}$, $\mathbf{g}$, $\mathbf{W}^{\text{noise}}$, and $\mathbf{H}^{\text{noise}}$. Once these variables are estimated, the MAP estimate, and equivalently the minimum mean squares estimate (MMSE), of the complex clean speech spectrum $\hat{x}^{\text{speech}}_{fn}$ is given by Wiener filtering:

$$\hat{x}^{\text{speech}}_{fn} = \frac{v^{\text{speech}}_{fn}}{v^{\text{speech}}_{fn} + v^{\text{noise}}_{fn}} x^{\text{mix}}_{fn}. \quad (9)$$

We can then reconstruct the time-domain speech estimate by taking the inverse STFT of $\hat{\mathbf{X}}^{\text{speech}}$.

Note that, the observation model in Eq 7 is different than the one defined in Eq 1. For this particular model, the update rules for $\mathbf{U}$ and $\mathbf{g}$ are slightly different than the ones defined in Section 3 and they can be achieved with a similar MM algorithm. The update rules for the SNMF model can be found in [24].

4.2. Experimental Setup

In our experiments, we use speech files from the TIMIT database and down-sample to 8 kHz. Signals are analyzed using the STFT with a sine window of length 320 samples and 75 % overlap for analysis and re-synthesis.

The parameters $\mathbf{A}^r$, $\mathbf{A}^e$, $\mathbf{B}$, and $\mathbf{W}^e$ of the NSFDS model are trained separately for male and female speech, each on 1000 utterances (about 50 minutes) from the TIMIT training set. To enforce a smooth filter component $\mathbf{V}^r$, we use as elementary filters $K^r = 10$ overlapping sine-shaped bandpass filters, uniformly distributed on the Mel-frequency scale (see $\mathbf{W}^r$ in Fig. 1). The number of elementary filters K^r should be small in order to prevent the filter part from capturing the excitation part. The number of phonemes in the training set is $I^r = 61$. We use $K^e = 300$ excitation profiles. For each mixture, we assume the gender is known and use the NSFDS model for that gender.

We evaluate the proposed method on mixtures of speech from the TIMIT test set with challenging non-stationary noise. The noise data were recorded in a car while driving in the Greater Boston area, and mainly include engine, road, blinker, wiper, rain, and city noises. For each of 40 utterances (20 female and 20 male), a noise signal is randomly selected and added to the speech at 3 different input signal-to-noise ratios (SNR), for a total of 120 mixtures.

4.3. Training Procedure

During training, we make use of reference information for the filter labels $\mathbf{h}^r$ and excitation labels $\mathbf{h}^e$, and keep those labels fixed to their reference values throughout the training process. For the filter labels $\mathbf{h}^r$, we use as reference labels the phoneme annotations provided with the TIMIT database. For the excitation labels $\mathbf{h}^e$, we allocate an excitation state to each unvoiced phoneme, and estimate the remaining (voiced) states by running a pitch estimator [25] on

Table 2. Evaluation results of the baseline methods and the proposed method.

Method	SNR = −20 dB			SNR = −10 dB			SNR = 0 dB		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
OM-LSA	0.75	6.53	3.81	10.09	16.93	12.15	18.88	26.60	21.06
VTS	4.92	10.22	5.86	11.96	19.84	14.06	19.01	27.55	21.43
i-VTS	4.34	13.00	7.27	11.01	25.17	14.79	18.25	27.11	21.95
SNMF	5.02	14.12	5.76	12.60	22.05	13.60	19.77	28.52	20.85
NDS	7.63	21.18	8.49	15.29	**28.10**	16.58	22.62	**33.96**	23.87
NSFDS (nd)	7.37	15.69	8.34	14.53	22.17	16.01	22.04	29.99	23.49
NSFDS (sl)	8.33	19.12	9.43	15.26	24.87	16.69	21.93	31.14	23.48
NSFDS	**9.18**	**21.27**	**10.10**	**16.17**	27.22	**17.45**	**22.66**	31.95	**23.99**

the speech training data and quantizing the obtained pitch estimates with the k-means algorithm.

By predefining the filter dictionary $\mathbf{W}^r$ to consist of smooth overlapping filters, we implicitly restrict the filter part $\mathbf{V}^r$ to capture the smooth envelope of the spectrum. However, since there is no explicit constraint on the excitation part $\mathbf{V}^e$, a good method for initializing the excitation dictionary $\mathbf{W}^e$ is key to ensure that $\mathbf{V}^e$ will capture only the pitch and other excitation modes. To initialize $\mathbf{W}^e$, we first compute the cepstrum $\mathbf{C} = \text{DCT}\{\log \mathbf{S}\}$, where DCT stands for the discrete cosine transform and $\mathbf{S}$ is the power spectrum of the training data. Eliminating the lower part of the cepstrum to remove the phoneme-related information, we define the *high-pass liftered spectrum*, $\mathbf{S}^{\text{high}} = \exp(\text{IDCT}\{\mathbf{C}^{\text{high}}\})$, where $c^{\text{high}}_{fn} = c_{fn}$ if $f > f_c$, and 0 otherwise, and f_c is a cut-off frequency. Finally, we initialize each column of $\mathbf{W}^e$ as the average of the corresponding columns of the liftered spectrum: $w^e_{fm} = (\sum_n [h^e_n = m] s^{\text{high}}_{fn}) / (\sum_n [h^e_n = m])$.

The variables $\mathbf{U}$ and $\mathbf{g}$ are initialized randomly under a uniform distribution. Once all the variables are initialized, we train the NSFDS model by using the update rules described in Section 3.

4.4. Testing Procedure

Initial conditions play an important role in alternating optimization methods. We here use the following initialization procedure.

We first run a simpler speech enhancement method, the optimally-modified log spectral amplitude estimator (OM-LSA) [1], on the noisy mixture. To initialize the pitch labels $\mathbf{h}^e$, we then run a pitch estimator on the OM-LSA output and initialize $\mathbf{h}^e$ accordingly. For the phoneme labels $\mathbf{h}^r$, we compute the low-pass liftered spectrum of the OM-LSA output and compare its columns with the columns of the low-pass liftered spectrum of the training data $\mathbf{S}^{\text{low}}$, where the low-pass liftered spectrum is defined similarly to its high-pass counterpart above, $\mathbf{S}^{\text{low}} = \exp(\text{IDCT}\{\mathbf{C}^{\text{low}}\})$, where $c^{\text{low}}_{fn} = c_{fn}$ if $f \leq f_c$, and 0 otherwise. Since reference phoneme labels for $\mathbf{S}^{\text{low}}$ are known, we can initialize $\mathbf{h}^r$ to the labels of the most similar columns of $\mathbf{S}^{\text{low}}$. The variables $\mathbf{U}$ and $\mathbf{g}$ are again initialized randomly under a uniform distribution.

After initializing the NSFDS model, we randomly initialize the SNMF noise model, run the noise model on the noise estimate of the OM-LSA algorithm until convergence and use these estimates as initial values for the noise model. Finally, we run our inference algorithm and obtain a clean speech estimate as described in Section 4.1.

4.5. Results

We measure the performance in terms of the signal to distortion ratio (SDR) signal to interference ratio (SIR), and signal to artifact ratio (SAR), using the BSS_{EVAL} toolbox v.3 [26]. We compare our method with state-of-the-art methods: OM-LSA, vector Taylor series (VTS) [9], indirect VTS (iVTS) [27], SNMF, and NDS. Among these methods, SNMF and NDS are gender dependent, i.e., they assume the gender is known. Similar to NSFDS, we combine SNMF and NDS speech models with an SNMF noise model as described in Section 4.1, where the noise models are initialized as in Section 4.4.

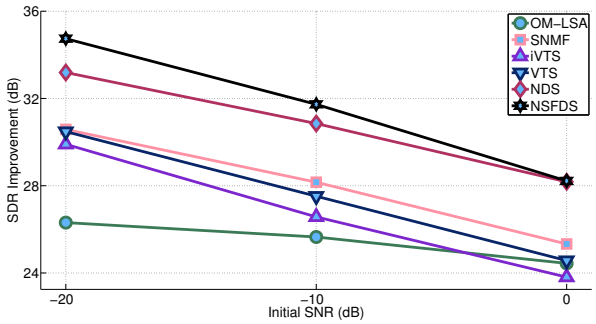

Fig. 3. SDR improvements for baseline and proposed models.

We also define two simpler versions of NSFDS to reveal the contributions of different parts of the model: in the first, NSFDS single-layer (sl), we discard the intermediate layer variables $\mathbf{B}$ and $\mathbf{U}$ and model the filter part exactly as the excitation part (see Eq. 2), training $\mathbf{W}^r$ as well; in the second, NSFDS no-dynamics (nd), we discard the temporal dependencies between $\mathbf{h}^r$, $\mathbf{h}^e$, and $\mathbf{g}$ and assume they are independent and identically distributed a priori.

For all models (including the baseline models), we investigate various parameter settings and report the best one in terms of SDR. The results are given in Table 2 and Fig. 3. Note that initial SNR is computed on parts where speech is present, while the SDR is computed on the whole mixtures, making initial SDR lower than initial SNR. The proposed NSFDS model outperforms all baseline methods in terms of SDR, with the improvement decreasing from -20 dB to 0 dB initial SNR. The results show that the usage of the intermediate layer and the dynamics each contribute to the performance, and the best performance is obtained with the full model. Note that the large baseline SDR improvements are due to the presence of easily-removable low-frequency stationary noise in the data. Informal subjective tests confirm that our method performs better than other methods; we invite the readers to check the audio samples available on our project webpage [23].

5. CONCLUSION

We presented a novel probabilistic model for speech enhancement following a source-filter approach in which the excitation and filter parts are modeled as non-negative dynamical systems. We presented convergence-guaranteed update rules for each latent factor. We evaluated our model on a challenging speech enhancement task involving non-stationary car noises, and showed that the proposed method outperforms the state-of-the-art in terms of objective measures.

6. REFERENCES

[1] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, pp. 113–116, 2002.

[2] 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.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2001, vol. 13, pp. 556–562.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[5] J. R. Hershey and M. Casey, "Audio-visual sound separation via hidden Markov models," in *NIPS*, vol. 2, pp. 1173–1180. MIT Press, 2002.

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

[7] U. Şimşekli, J. Le Roux, and J. R. Hershey, "Hierarchical and coupled non-negative dynamical systems with application to audio modeling," in *WASPAA*, 2013.

[8] J. R. Hershey, S. J. Rennie, and J. Le Roux, "Factorial models for noise robust speech recognition," in *Techniques for Noise Robustness in Automatic Speech Recognition*, T. Virtanen, R. Singh, and B. Raj, Eds., chapter 12. Wiley, 2012.

[9] T. Kristjansson and J. R. Hershey, "High resolution signal reconstruction," in *ASRU*, 2003.

[10] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in neural information processing systems*, pp. 758–764, 2001.

[11] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, 1998.

[12] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, 1998.

[13] G. Mysore and M. Sahani, "Variational inference in non-negative factorial hidden Markov models for efficient audio source separation," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, J. Langford and J. Pineau, Eds., New York, NY, USA, July 2012, ICML '12, pp. 1887–1894, Omnipress.

[14] G. Fant, *Acoustic theory of speech production*, Mouton, The Hague, 1970.

[15] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization," in *Advances in models for acoustic processing, neural information processing systems workshop*. Citeseer, 2006.

[16] D. Fitzgerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Inteligence and Neuroscience*, vol. 2008, 2008.

[17] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," in *ICASSP*, 2010.

[18] J. L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 564–575, 2010.

[19] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–filter-based single-channel speech separation using pitch information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, 2011.

[20] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation.," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1118–1133, 2012.

[21] 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, pp. 793–830, 2009.

[22] M. Nakano, H. Kameoka, J. Le Roux, Y. Kitano, N. Ono, and S. Sagayama, "Convergence-guaranteed multiplicative algorithms for non-negative matrix factorization with beta-divergence," in *MLSP*, Aug. 2010.

[23] U. Şimşekli, J. Le Roux, and J. R. Hershey, "NSFDS project webpage: Supplementary document and sound samples," `http://www.merl.com/demos/speech-enhancement-NSFDS`, 2014, [Online].

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

[25] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[26] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *ICA*, 2007, pp. 552–559.

[27] J. Le Roux and J. R. Hershey, "Indirect model-based speech enhancement," in *ICASSP*, Mar. 2012.

NON-NEGATIVE SOURCE-FILTER DYNAMICAL SYSTEM FOR SPEECH ENHANCEMENT SUPPLEMENTARY MATERIAL

Umut Şimşekli,[1] *Jonathan Le Roux,*[2] *John R. Hershey*[2]

[1]Boğaziçi University, Dept. of Computer Engineering, 34342, Bebek, İstanbul, Turkey
[2]Mitsubishi Electric Research Laboratories (MERL), 201 Broadway, Cambridge, MA 02139, USA
umut.simsekli@boun.edu.tr, {leroux, hershey}@merl.com

1. INFERENCE

In this section, we present convergence-guaranteed update rules for maximum a-posteriori (MAP) estimation in the proposed model. In particular, we use the majorization-minimization (MM) algorithm which monotonically decreases the intractable MAP objective function by minimizing a tractable upper-bound constructed at each iteration. This algorithm is a block-coordinate descent algorithm which performs alternating updates of each latent factor given its current value and the other factors. The MM algorithm yields the following updates for $\mathbf{B}$ and $\mathbf{W}^e$:

$$b_{ki} \leftarrow \frac{\beta \sum_n [h_n^r = i] u_{kn}}{\alpha \sum_n [h_n^r = i]}, \quad w_{fm}^e \leftarrow \frac{\sum_n [h_n^r = m] \frac{v_{fn}}{g_n v_{fn}^r}}{\sum_n [h_n^e = m]} \tag{1}$$

The update for the transition matrix $\mathbf{A}^r$ is as follows:

$$a_{ij}^r \leftarrow \frac{\sum_n [h_n^r = i][h_{n-1}^r = j]}{\sum_n [h_n^r = i]} \tag{2}$$

where the update for $\mathbf{A}^e$ is identical to Equation 2 up to replacing variables h_n^r with h_n^e.

The updates of $\mathbf{U}$ and $\mathbf{g}$ involve finding roots of second order polynomials. The corresponding equations are given in Table 1. Besides, given all other variables, the optimal $\mathbf{h}^r$ and $\mathbf{h}^e$ can be computed via Viterbi algorithm at each iteration.

Table 1. Update rules for $\mathbf{U}$ and $\mathbf{g}$ for clean speech. The factors can be updated at each iteration to the value $\frac{\sqrt{b^2-4ac}-b}{2a}$ where each factor has different a, b, and c values. Here, we define $\hat{v}_{fn} = g_n v_{fn}^r v_{fn}^e$.

	a	b	c
u_{kn}	$\sum_f \frac{w_{fk}^r}{v_{fn}^r} + \frac{\beta}{\prod_i b_{ki}^{[h_n^r=i]}}$	$1-\alpha$	$-u_{kn}^2 \sum_f \frac{v_{fn}}{g_n v_{fn}^e \hat{v}_{fn}^2} w_{fk}^r$
g_n $(n=1)$	$(F+\phi)^2$	0	$-\left[\sum_f \frac{v_{fn}}{v_{fn}^r v_{fn}^e} + \psi g_{n+1}\right]^2$
g_n $(1<n<N)$	$\frac{\psi}{g_{n-1}}$	$F+1$	$-\left[\sum_f \frac{v_{fn}}{v_{fn}^r v_{fn}^e} + \psi \frac{g_n}{g_{n+1}}\right]$
g_n $(n=N)$	$\frac{\psi}{g_{n-1}}$	$F+1-\phi$	$-\left[\sum_f \frac{v_{fn}}{v_{fn}^r v_{fn}^e}\right]$

Table 2. Update rules for $\mathbf{U}$ and $\mathbf{g}$ for noisy mixture. The factors can be updated at each iteration to the value $\frac{\sqrt{b^2-4ac}-b}{2a}$ where each factor has different a, b, and c values. Here, we define $\hat{v}_{fn} = v_{fn}^{\text{speech}} + v_{fn}^{\text{noise}}$, $\bar{v}_{fnk} = g_n w_{fk}^r v_{fn}^e / \hat{v}_{fn}$, and $v_{fn}^{re} = v_{fn}^r v_{fn}^e$.

	a	b	c
u_{kn}	$\sum_f \bar{v}_{fnk} + \frac{\beta}{\prod_i b_{ki}^{[h_n^r=i]}}$	$1-\alpha$	$-u_{kn}^2 \sum_f \frac{v_{fn} \bar{v}_{fnk}}{\hat{v}_{fn}}$
g_n $(n=1)$	$\sum_f \frac{v_{fn}^{re}}{\hat{v}_{fn}}$	ϕ	$-\left[g_n^2 \sum_f \frac{v_{fn} v_{fn}^{re}}{\hat{v}_{fn}^2} + \psi g_{n+1}\right]$
g_n $(1<n<N)$	$\sum_f \frac{v_{fn}^{re}}{\hat{v}_{fn}} + \frac{\psi}{g_{n-1}}$	1	$-\left[g_n^2 \sum_f \frac{v_{fn} v_{fn}^{re}}{\hat{v}_{fn}^2} + \psi g_{n+1}\right]$
g_n $(n=N)$	$\sum_f \frac{v_{fn}^{re}}{\hat{v}_{fn}} + \frac{\psi}{g_{n-1}}$	$1-\phi$	$-\left[g_n^2 \sum_f \frac{v_{fn} v_{fn}^{re}}{\hat{v}_{fn}^2}\right]$