A STATISTICAL APPROACH TO SEMI-SUPERVISED SPEECH ENHANCEMENT WITH LOW-ORDER NON-NEGATIVE MATRIX FACTORIZATION

Shoaib Mohammed and Ivan Tashev

Microsoft Research, One Microsoft Way, Redmond WA 98052

ABSTRACT

Compared to generic source separation, NMF for speech enhancement is relatively underexplored. When applied to the latter problem, NMF is bereft of performance consistency (across runs and data samples), *esp.* with small-sized dictionaries. This limitation raises the need for higher-order representations, leading to increased computational costs. In this paper, we propose a statistical-estimation technique that attempts to bridge this gap. Our approach combines multiple low-order NMF decompositions of noisy speech to increase the overall enhancement performance. We show PESQ improvements of up to 0.24 beyond what is achievable by a single NMF parametrization and, at iso-performance levels, major reductions in computational cost.

Index Terms— Semi-supervised speech enhancement, spectral estimation, non-negative matrix factorization

1. INTRODUCTION

Sparsity is an important property of data that is exploited in a variety of signal-processing problems [1, 2]. In the context of non-negative matrix factorization (NMF), sparsity allows us to control the uniqueness of signal representation. Consider, for instance, the case of speech data. NMF enables us to factorize amplitude spectrograms $\mathbf{V}_{M\times T}$ into a product of dictionary atoms $\mathbf{W}_{M\times K}$ and activations $\mathbf{H}_{K\times T}$. The interesting cases of $K \lessgtr M$ (denoting under- and over-complete representations) have many possible parametrizations. In such settings, imposing constraints of sparsity through a regularization term limits the non-uniqueness of NMF to a certain degree [3]. Unfortunately, this is often not enough. Non-convex problems like these are solved through iterative updates that can only guarantee local minima [4]. Thus the quality of NMF results depend heavily on the initialization strategy used.

We performed an empirical study to assess the factorization performance of NMF. Fig. 1 shows the perceptual quality (measured by PESQ [5]) of clean speech in the TIMIT data corpus [6] when represented by many equisized NMF dictionaries that are initialized with distinct random seeds [4]. From the figure, we make two important observations: (1) mean representation accuracy is poor and less predictable across speaker files (shown as error bars in the figure), *esp.* at small values of K, and (2) there is also variation across equisized dictionaries (shown as a histogram in the inset) of the same speaker file, which is again amplified at smaller values of K. Thus, we conclude that under-complete representations of clean speech with NMF decomposition ($K \ll M$, $M = 256$ in our case) are less-robust, necessitating the use of additional basis functions (higher-order dictionaries) to accurately represent speech. As shown in the left side of Table 1, these effects persist across other initialization strategies: PCA [7, 8], KMeans (KM) [9, 10], (EX)NMF [11]. Thus, our observations raise an important issue of performance *vs.* efficiency in NMF. The right side of Table 1 shows that the incurred processing and memory overheads, size and bandwidth (BW), associated with high-order NMF decompositions can be substantial (results in the table are estimated for processing 1 sec. of speech).

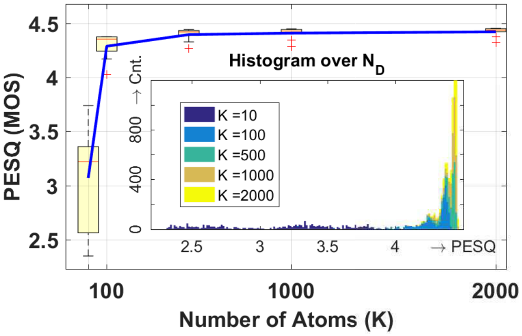

Fig. 1: Low-order NMF is non-robust: variation across files (error bars) and dictionary initializations (inset histogram).

In this paper, we attempt to achieve high representation accuracy with low-order NMF decompositions (low computational costs). Our approach is to use a statistical-estimation technique that exploits the diversity in NMF factorizations across multiple local equisized dictionaries. To appreciate the insight, consider the example spectra shown in Fig. 2 obtained by factorizing speech signals *via* 40 NMF dictionaries, initialized with distinct random seeds. The variation in the frequency components is consistent with the observations made in Fig. 1. Our proposed approach, abbreviated henceforth as NMFSE (NMF with spectral estimation), intercepts such spectral estimates in the semi-supervised speech enhancement process, models each time-frequency bin as a random variable, invokes the spatio-temporal continuity prop-

Table 1: Variation across init. strategies and NMF costs.

K ↓		**Rand.**	**PCA**	**KM**	**EX**	**#OPs**	**Mem.**	**BW**
10	PESQ	3.20	3.22	3.26	2.96	124 MFLOPs	0.88 MB	7.62 Mbps
	std.dev.	0.18	0	0.04	0.25			
1000	PESQ	4.43	–	4.43	4.07	11.43 GFLOPs	10.98 MB	149 Mbps
	std.dev.	0.01	–	0.01	0.03			

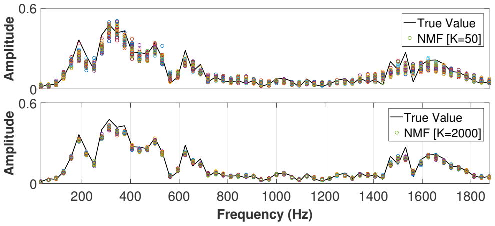

Fig. 2: Variation in spectra across equisized NMF dictionaries

erties of speech signals to build a robust prior, and eventually estimates the most likely value of the spectrum. Through several experiments, we show that NMFSE gives consistent speech-enhancement performance that is better than a single NMF, while incurring little computational overheads.

2. BACKGROUND AND RELATED WORK

When used for speech enhancement, NMF allows us to infer latent structures (speech is sparse and noise is not) in noisy speech signals by factorizing their amplitude spectrograms **V** into a linear combination of basis functions **W** that define a convex cone as follows:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} = \underset{\bar{\mathbf{W}},\bar{\mathbf{H}}}{\operatorname{argmin}} \left[\mathrm{D}\left(\mathbf{V}\|\bar{\mathbf{W}}\bar{\mathbf{H}}\right) + \nu \left\|\bar{\mathbf{H}}\right\|_1 \right], \tag{1}$$

where ν controls the sparsity weight and D represents one of many possible divergence metrics [4]. Unfortunately, Eq. (1) has no closed form solution and multiplicative updates are often used to get the best approximation [3, 4, 12, 13]. All of the update algorithms start out with an initial seed for **W** and **H**, and continue to refine the estimates iteratively until they reach the desired level of error convergence.

Although NMF has been used for speech enhancement [14], past studies have shown that the quality of NMF factorizations are sensitive to initializations [8, 15]. There have also been efforts to develop better initialization strategies [7, 8]. Unfortunately, because globally optimal solutions cannot be guaranteed, it is likely that some random initialization will beat the best proposed strategy. This is also perhaps the reason behind why the latter is the most widely adopted strategy in the literature [3, 4, 14]. Further, papers that propose alternative initializations do not dwell on performance variation across different runs (*e.g.*, K-means initialization is driven by the choice of the first centroid) [9, 10]. Through a thorough experimental study, we shed more light on the variability aspect of NMF, addressing a limitation of the literature.

There has been prior work that makes use of multiple dictionaries in NMF [16–18]. These methods learn dictionaries with subsets of data or block-sparse constraints [19–21]. Unfortunately, they also suffer from the initialization problem leading to less predictable (widely varying) performance. Our methodology is similar in spirit to other post-processing techniques like spectral smoothing [22] and ada boosting [23, 24] as well as analyses constrained by the spatio-temporal dynamics of speech [12, 25].

3. PROPOSED APPROACH

Fig. 3 shows an overview of the proposed approach. Our goal is to separate speech signals $x_s(t)$ from noisy recordings $y(t)$ in a single channel. First, we use clean speech data to train N_D equisized low-order dictionaries (small number of atoms or basis functions), $\mathbf{W}_{si}$, $i = 1, \ldots, N_D$, offline with distinct random initializations. At runtime, we use T time windows to compute the $M \times T$ dimensional short-time Fourier transform (STFT), $\mathbf{Y}(t, f)$, of the noisy signal $y(t)$. Assuming independence and linear mixing of speech and noise, we fix the N_D speech dictionaries and use NMF to factorize the amplitude spectrogram, $\mathbf{V}$, of $\mathbf{Y}(t, f)$ as follows:

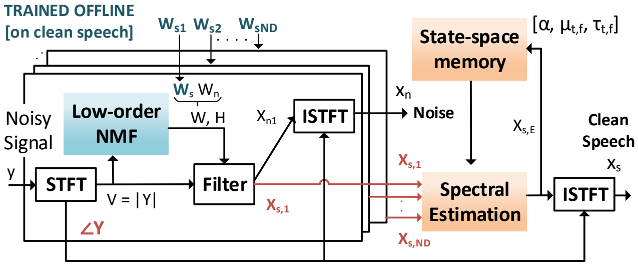

Fig. 3: Block diagram of the proposed NMFSE approach.

$$\mathbf{V} \approx [\mathbf{W}_{si}, \mathbf{W}_{ni}]\,[\mathbf{H}_{si}, \mathbf{H}_{ni}] \tag{2}$$

where noise dictionaries, $\mathbf{W}_{ni}$ and activations, $\mathbf{H}_{si}$ and $\mathbf{H}_{ni}$ are determined using the sparse NMF formulation [Eq. (1)]. For each factorization, we reconstruct the speech spectrogram, $\mathbf{X}_{si}$, *via* a generalized Wiener-filtering approach:

$$\mathbf{X}_{si} = \frac{(\mathbf{W}_{si}\mathbf{H}_{si})^p}{(\mathbf{W}_{si}\mathbf{H}_{si})^p + (\mathbf{W}_{ni}\mathbf{H}_{ni})^p} \tag{3}$$

where p (typically in the range 1-2) defines the smoothness of the masking function. The N_D spectrograms, comprising time-frequency components X^{tf}_{si}, $t = 1, \ldots, T$, $f = 1, \ldots, M$, are then processed by the spectral estimation technique described below.

3.1. Spectral Estimation

To determine the final amplitude spectrogram of the clean speech signal, we consider NMF-estimated values in the time-frequency bins to be independent and identically distributed Gaussian random variables, $X^{tf}_{si} \sim \mathcal{N}(\mu, \tau^2_s)$, whose N_D sample values, given by the reconstructed signals $X^{tf}_{s1}, \ldots, X^{tf}_{sN_D}$, are used to determine μ and τ^2_s. We then proceed to obtain the maximum *a postériori* probability (MAP) estimate of X^{tf}_{SE} as follows:

$$X^{tf}_{SE,MAP} = \underset{X^{tf}_{SE}}{\operatorname{argmax}}\, P(X^{tf}_{SE}|X^{tf}_{si}) = \underset{X^{tf}_{SE}}{\operatorname{argmax}}\, P(X^{tf}_{si}|X^{tf}_{SE})P(X^{tf}_{SE}) \tag{4}$$

where $P(X^{tf}|X^{tf}_{SE})$ is the likelihood of the time-frequency component X^{tf} and $P(X^{tf}_{SE})$ is the prior distribution of X^{tf}_{SE}.

Further, we assume $P(X^{tf}_{SE}) = \mathcal{N}(\mu_0, \tau^2_m)$, a conjugate prior. In order to determine μ_0 and τ^2_m, we leverage the spatio-temporal continuity property of speech signals. Fig. 4 illustrates our approach. To the right of the figure is a time-frequency slice of the spectrogram of an example clean-speech signal. To determine μ_0 for a sample of interest X^{tf}_{SE}, we use a weighted sum of the sample means, $\mu_{t,f-1}$ and $\mu_{t-1,f}$,

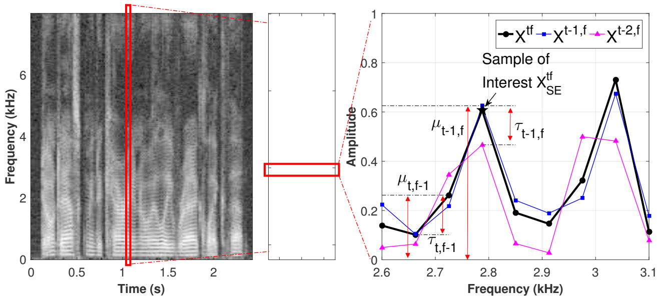

Fig. 4: Illustration of the various components used in the computation of the prior mean and variance.

from the preceeding spatio-temporal bins, $X_{si}^{t,f-1}$ and $X_{si}^{t-1,f}$, respectively, as follows:

$$\mu_0 = \alpha\,\mu_{t,f-1} + (1-\alpha)\mu_{t-1,f} \tag{5}$$

where α is a parameter that controls the spatio-temporal (ST) scale. Similarly, we determine the prior variance τ_m^2 using a weighted sum of the gradients in the sample mean as shown below:

$$\begin{aligned}\tau_m^2 &= \left[\alpha\left(\frac{\partial\mu_{t,f}}{\partial f}\right)_{f-1} + (1-\alpha)\left(\frac{\partial\mu_{t,f}}{\partial t}\right)_{t-1}\right]^2\\ &\approx \left[\alpha\left(\mu_{t,f-1} - \mu_{t,f-2}\right) + (1-\alpha)\left(\mu_{t-1,f} - \mu_{t-2,f}\right)\right]^2 \end{aligned}\tag{6}$$

Thus, solving Eq. (4), we obtain the non-negative MAP estimate of each time-frequency bin as follows:

$$X_{SE,MAP}^{tf} = \max\left[0, \frac{M\tau_m^2}{M\tau_m^2+\tau_s^2}\left(\frac{1}{M}\sum_{j=1}^{M} X_{sj}^{tf}\right) + \frac{\tau_s^2}{M\tau_m^2+\tau_s^2}\mu_0\right] \tag{7}$$

where τ_s^2 is the posterior sample variance obtained from N_D NMF estimates of X_{SE}^{tf}. The joint MAP estimate of the spectral components, X_{SE}, is fused with phase information from the initial synthesis to obtain the clean speech spectrogram, which is then passed through an inverse STFT block to get the final clean speech signal $x_s(t)$. A summary of the end-to-end process is presented in Algorithm 1.

Algorithm 1 NMFSE with spatio-temporal continuity priors

```
Input: Noisy signal $y(t)$, Speech corpus $\mathbf{D}_{tr}$, ST scale $\alpha$
Output: Clean speech signal $x(t)$
Train dictionaries: $\mathbf{W}_{si} \leftarrow \text{NMF}(\mathbf{D}_{tr})$, $i = 1, \ldots, N_D$
initialize $\mathbf{V}_{M\times T} = \left|\text{STFT}[y(t)]\right|$, $\Phi = \angle\text{STFT}[y(t)]$
for i :=1 to $N_D$ do                                  // Over all dictionaries
    $\mathbf{X}_{si} \leftarrow \text{NMF}[\mathbf{V}_{M\times T}, \mathbf{W}_{si}, K]$
end for
for [t, f] :=1 to [T, M] do                            // Over all ST bins
    $X_{SE}^{tf} \leftarrow X_{SE,MAP}^{tf}$ from Eq.(7)
end for
$x(t) = \text{ISTFT}\left[\mathbf{X}_{SE}\cdot\mathbf{e}^{j\Phi}\right]$
```

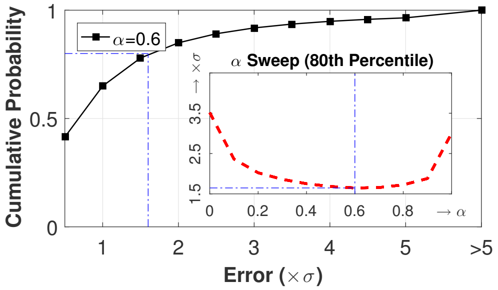

Fig. 5: With α=0.6, most data samples lie within 1.7×σ.

4. EXPERIMENTAL RESULTS

Our main goal is to show the benefits of using the proposed NMFSE algorithm for semi-supervised speech enhancement with low-order NMF dictionaries. In this section, we describe the data we use, our experimental setup, and results that validate the proposed approach.

4.1. Data and Setup

To capture the phonetical richness of speech, we used data from the TIMIT corpus [6]. We used 80% of the corpus to train 18 sets of 100 speech dictionaries offline using combinations of K (= 10, 50, 100, 500, 1000, 2000) atoms and ν (= 0.1, 0.25, 1) sparsity levels. We randomly initialized every dictionary and used sparse NMF decomposition with KL-divergence [3, 4]. We pre-processed the signals using Hann windows with 50% overlap and FFT sizes of 512 points.

To limit experimentation time, we randomly selected 10 speech files among the remaining 20% of the corpus as test instances and added noise segments to them – at matched sampling rates of 16kHz – from the NOISEX-92 corpus at four input signal-to-noise ratio (SNR) levels: -3, 0, 6, and 12 dB [26]. We fixed the post-filter masking parameter p to be 1. In order to assess the perceptual quality of speech, we used the PESQ metric defined by the ITU-T [5]. We also measured the SNR and signal-to-distortion ratio (SDR) of the speech signals [27]. As a baseline, we used a classic noise suppressor that is a generalization of the decision-directed approach, first defined in [28]. We present our measurement results below.

4.2. Evaluation

First, we find the ST scale factor α that minimizes the error in the MAP estimates of TIMIT clean speech [Eq. (7)]. We did

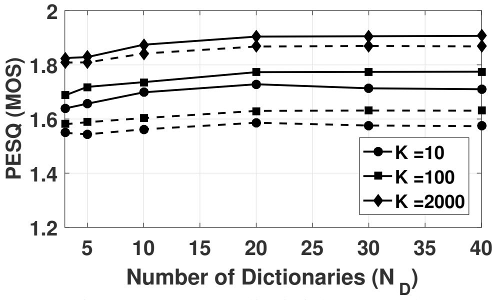

Fig. 6: PESQ improvements diminish beyond N_D=20. Solid lines are for NMFSE and dashed lines are for NMF.

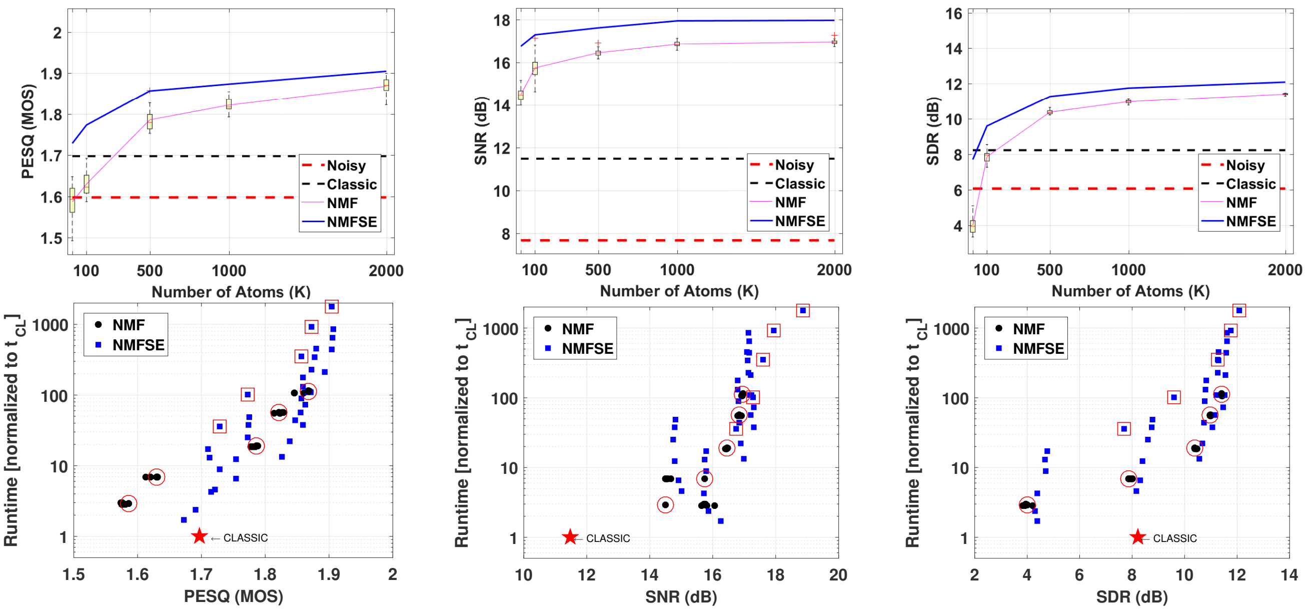

Fig. 7: **Top:** NMFSE improves NMF PESQ, SNR and SDR by up to 0.24, 2.76 dB and 4.37 dB, respectively. **Bottom:** runtimes with different values of N_D; points corresponding to the upper figures (N_D=20) are highlighted with big markers.

a sweep of alpha and found that $\alpha = 0.6$ gives us the best ST scale trade-off [see inset in Fig. 5]. We chose this value for the enhancement experiments. Fig. 5 also shows that at $\alpha = 0.6$, the 80^{th} percentile of the cummulative MAP error across our training dataset falls within 1.7 times the estimated standard deviation σ. The other parameter to determine for NMFSE is the optimal number of dictionaries to use N_D. Through an empirical evaluation of PESQ *vs.* N_D [the case of SNR = 6 dB is shown in Fig. 6], we found that the benefits of combining dictionaries with NMFSE diminish beyond N_D = 15-20. Thus, we chose $N_D = 20$ for the rest of the experiments.

Fig. 7 (top part) shows mean enhancement results (mean across all test files) at input SNR level = 6 dB and $\nu = 0.25$. From the figure, we see that at small values of K, NMF performs poorly compared to the classic noise suppressor. There is also a substantial variation in the enhancement performance across runs (shown as error bars), depending on the dictionary $\mathbf{W}_{si}$ that we use. However, performance of NMF improves when K increases. NMFSE betters these results. In the average case, it improves NMF PESQ by up to 0.14, SNR by 2.26 dB and SDR by 3.71 dB. In the best case, the corresponding improvements are up to 0.24 (substantial), 2.76 dB and 4.37 dB. The benefits of NMFSE are pronounced at smaller values of K (higher and consistent performance).

Fig. 7 (bottom part) shows that at any given value of K, NMFSE incurs higher computational costs (N_D=20) when compared to a single NMF (runtimes are normalized with respect to that of the classic noise suppressor, t_{cl}). However, at iso-performance levels, NMFSE costs can be substantially lower since it can achieve better performance with small values of K. The improvements in PESQ due to NMFSE are shown across different input SNR and ν levels in Table 2. The highlighted rows show notable improvements and validate the fact that, when operated at the pareto-optimal points in Fig. 7, NMFSE can indeed improve enhancement performance with lower computational costs than NMF. For instance, at SNR = 0 dB, NMFSE increases PESQ of the noisy signal by 0.51 at a much lower computational cost (32×) than NMF (0.39 increase at 53× cost).

5. CONCLUSIONS

NMFSE is a statistical post-processing technique that improves speech enhancement performance with small-sized NMF dictionaries. A key benefit of using NMFSE is that it produces consistent improvements in PESQ, SNR and MSE (low variation across files and runs), and thus alleviates the uncertainty associated with different initialization points in NMF. A proper design-space exploration enables us to reap these benefits at lower computational costs than NMF.

Table 2: NMFSE improves performance across different SNR and ν values (N_D=20) at lower computational costs than NMF.

	ΔPESQ @ input SNR=-3dB				**ΔPESQ @ input SNR= 0dB**				**ΔPESQ @ input SNR=12dB**			
	K=10		**K=1000**		**K=10**		**K=1000**		**K=10**		**K=1000**	
$\nu\downarrow$	NMF	NMFSE	NMF	NMFSE	NMF	NMFSE	NMF	NMFSE	NMF	NMFSE	NMF	NMFSE
0.1	0.10	0.11	0.11	0.19	0.13	0.16	0.16	0.21	0.07	0.16	0.14	0.28
0.25	0.16	0.24	0.28	0.30	0.20	0.29	0.36	0.41	**0.12**	**0.37**	**0.32**	0.38
1.0	**0.25**	**0.33**	**0.47**	0.54	**0.31**	**0.51**	**0.39**	0.58	0.17	0.51	0.44	0.63
Norm. RT →	**1.3×**	**17×**	**28×**	487×	**2.7×**	**32×**	**53×**	899×	**2.7×**	**33.3×**	**54.8×**	924×

6. REFERENCES

[1] R. G. Baraniuk, E. Candes, M. Elad, and L. Yu, "Applications of sparse representation and compressive sensing," in *Proc. IEEE*, Jul. 2010, pp. 906–909.

[2] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, May 2015.

[3] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Machine Learning Research*, vol. 5, pp. 1457–1469, Nov. 2004.

[4] D. D. Lee and S. H. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*, Dec. 2001, pp. 556–562.

[5] C. Rusu and B. Dumitrescu, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, May 2001, pp. 749–752.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Philadelphia, Jan. 1993.

[7] Z. Zheng, J. Yang, and Y. Zhu, "Initialization enhancer for non-negative matrix factorization," *Appl. Aritificial Intelligence*, vol. 20, no. 1, pp. 101–110, Apr. 2006.

[8] Y.-D. Kim and S. Choi, "A method of initialization for non-negative matrix factorization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Apr. 2007, pp. 533–537.

[9] Y. Xue, C. S. Tong, Y. Chen, and W.-S. Chen, "Clustering-based initialization for non-negative matrix factorization," *Applied Mathematics and Computation*, vol. 205, no. 2, pp. 525–536, Nov. 2008.

[10] S. Wild, J. Curry, and A. Dougherty, "Improving non-negative matrix factorizations through structured initialization," *J. Pattern Recognition*, vol. 37, no. 11, pp. 2217–2232, Nov. 2004.

[11] B. King and L. Atlas, "Single-channel source separation using simplified-training complex matrix factorization," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Apr. 2010, pp. 4206–4209.

[12] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," in *IEEE Trans. Audio, Speech and Lang. Proc.*, Mar. 2007, vol. 15, pp. 1066–1074.

[13] J. L. Roux, F. Weninger, and J. R. Hershey, "Sparse NMF – half-baked or well done?," Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., No. TR2015-023, Jun. 2015.

[14] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech and Language Processing*, vol. 21, no. 10, pp. 2140–2151, Jun. 2013.

[15] E. M. Grais and H. Erdogan, "Initialization of nonnegative matrix factorization dictionaries for single channel source separation," in *Int. Conf. Signal Proc. and Communications Applications*, May 2013, pp. 1–4.

[16] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Signal Processing Let.*, vol. 22, no. 3, pp. 293–297, Mar. 2015.

[17] J. L. Roux, S. Watanabe, and J. R. Hershey, "Ensemble learning for speech enhancement," in *Wkshp. Appl. Signal Proc. Audio and Acoustics*, Oct. 2013, pp. 1–4.

[18] D. Greene, G. Cagney, N. Krogan, and P. Cunningham, "Ensemble non-negative matrix factorization methods for clustering protein–protein interactions," *Bioinformatics*, vol. 24, no. 15, pp. 1722–1728, Aug. 2008.

[19] G. J. Mysore, "A block sparsity approach to multiple dictionary learning for audio modeling," in *Proc. Int. Wkshp. Sparsity, Dictionaries, Projections, Machine Learning and Signal Processing*, Jun. 2012.

[20] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, May. 2013, pp. 141–145.

[21] X. Jaureguiberry, P. Leveau, S. Maller, and J. J. Burred, "Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Apr. 2011, pp. 5–8.

[22] E. M. Grais and H. Erdogan, "Spectro-temporal post-smoothing in NMF based single-channel source separation," in *European Signal Processing Conf.*, Aug. 2012, pp. 584–588.

[23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[24] D. Li, Z. Su, H. Li, and X. Luo, "Boosting accuracy of attribute prediction via SVD and NMF of instance-attribute matrix," in *Pacific Rim Conf. Multimedia*, Dec. 2015, pp. 466–476.

[25] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, May 2011, pp. 17–20.

[26] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, Jun. 1993.

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

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