SPEAKER AND NOISE INDEPENDENT ONLINE SINGLE-CHANNEL SPEECH ENHANCEMENT

François G. Germain

Center for Computer Research in Music and Acoustics
Stanford University, Stanford, CA, USA
fgermain@stanford.edu

Gautham J. Mysore

Adobe Research
San Francisco, CA, USA
gmysore@adobe.com

ABSTRACT

Desirable properties of real-world speech enhancement methods include online operation, single-channel operation, operation in the presence of a variety of noise types including non-stationary noise, and no requirement for isolated training examples of the specific speaker and noise type at hand. Methods in the literature typically possess only a subset of these properties. Source separation methods particularly rarely simultaneously possess the first and last properties. We extend universal speech model-based speech enhancement to adaptively learn a noise model in an online fashion. We learn a model from a general corpus of speech in place of speaker-dependent training examples before deployment. This setup provides all of these desirable properties, making it easy to deploy in real-world systems without the need to provide additional training examples, while explicitly modeling speech. Our experimental results show that our method achieves the same performance as in the case in which speaker-dependent training data is available.

Index Terms— online speech enhancement, non-negative matrix factorization, universal speech models

1. INTRODUCTION

Speech enhancement refers to the task of improving the quality and/or the intelligibility of speech, often to compensate for the degradation introduced by the presence of a background noise signal (in this paper, we refer to noise as any interfering source that is not speech). Applications of this task are numerous and range from improved intelligibility and fidelity for hearing aids and mobile phones to front ends for automatic speech recognition. For many such real-world applications, these methods need to operate online and in real time.

Speaker and noise independent speech enhancement refers to performing this task on a mixture of speech from any speaker with any type of noise without the availability of training data from the given speaker or noise type. This poses serious challenges due to the need for additional information about the sources. In [1], we find that traditional speech enhancement methods can be categorized into four classes: spectral subtraction [2], Wiener filtering [3], statistical-based model methods [4] and subspace methods [5]. More recently, methods inspired by source separation methods and based on non-negative matrix factorization (NMF) [6, 7] or probabilistic latent component analysis (PLCA) [8] have been applied successfully to the task of online speech enhancement. All these methods present the limitation that they require the estimation of an associated noise model (e.g., spectrum, subspace), making them noise-dependent, since their performance relies heavily on the quality of that estimation. Methods such as Wiener filtering or spectral subtraction can be somewhat more robust than the others to unseen noise when the noise is stationary, but their performance decreases drastically in the presence of non-stationary noise. Some existing methods rely on available spatial information when multiple channels are available (e.g., in stereo mixtures) [9, 10]. Other methods (e.g., [7]) rely on a learned model representing the speech structure of the speaker in order to discriminate speech from noise sources. Those methods are then speaker-dependent, meaning dependent on the availability of training data associated with that specific speaker.

In many practical scenarios, we have to deal with the situation in which only a single-channel mixture is available and we have no prior knowledge of either the speech or noise. The above-mentioned methods are no longer applicable in such scenarios, either due to the absence of information on the sources, or of spatial information. Additionally, some of them prove to be unsuitable for online implementation. By making simple assumptions on the structure of the speech and the noise signals, we can derive methods leading to convincing speech enhancement. A recent method [11] uses the REPET-SIM method [12] which assumes that the background noise has a dense and low-ranked (i.e. repetitive) structure while the speech has a sparse and time-varying structure. This is an offline method, but its low computational cost allows it to run in real-time on a sliding buffer advancing on the spectrogram one frame at a time, effectively making it online.

In [13], an NMF-based method was presented that allows for speaker and noise independent offline speech enhancement. Though the method is based on a semi-supervised source separation NMF method [14], the method does not rely on the availability of speaker-dependent data to train the speech model. Instead, a collection of models is learned from a set of independent speakers and concatenated in order to form a so-called *universal speech model* (USM). The method then iteratively selects the closest speaker(s) matching the speech signal in the mixture in order to perform speaker independent separation. In addition to performing as well on the separation task as the case in which speaker-dependent training data is available, this framework was also successfully applied to the task of speaker and noise independent voice activity detection [15].

In this paper, we present in Section 2 our method that adapts the online semi-supervised PLCA-based method presented in [8] to incorporate the USM framework. This effectively removes the need to obtain speaker-dependent training data, resulting in a speaker and noise independent online method. In Section 3, we show the results of a series of experiments demonstrating that this technique performs as well as its speaker-dependent counterpart, while outperforming an existing speaker and noise independent method, as well as traditional speech enhancement techniques.

2. ALGORITHMS

2.1. Background

NMF-based source separation methods [14, 16] take advantage of the non-negative nature of the magnitude spectrogram $\mathbf{S}$ of a signal to approximate it as $\mathbf{S} \approx \mathbf{WH}$. In this formulation, the frame $\mathbf{s}_t$ at time t can be expressed as:

$$\mathbf{s}_t \approx \sum_k h_{k,t}\mathbf{w}_k \tag{1}$$

In many audio processing methods, as in NMF, it is generally acceptable to approximate the spectrogram as the linear superposition of the spectral features associated with the audio events present in the mixture. Following that, we can interpret this equation as approximating the frame $\mathbf{s}_t$ at time t as the superposition of the spectral features $\mathbf{w}_k$ weighted by the activation $h_{k,t}$, and the whole magnitude spectrogram $\mathbf{S}$ as the matrix multiplication $\mathbf{WH}$ with $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_K]$ and $\mathbf{H} = [h_{k,t}]$. The factorization is obtained by solving the optimization problem:

$$\underset{\mathbf{W},\mathbf{H}\geq 0}{\text{minimize}}\ D(\mathbf{S}||\mathbf{WH}) \tag{2}$$

with D the cost function, subject to the constraint that $\mathbf{W}$ and $\mathbf{H}$ are non-negative matrices. Here, we use the generalized Kullback-Leibler (KL) divergence as cost function as it is commonly used in source separation. It is defined as:

$$D(\mathbf{S}||\hat{\mathbf{S}}) = \sum_{f,t} s_{f,t}\log\left(s_{f,t}/\hat{s}_{f,t}\right) - s_{f,t} + \hat{s}_{f,t} \tag{3}$$

where $\mathbf{S} = [s_{f,t}]$ and $\hat{\mathbf{S}} = \mathbf{WH} = [\hat{s}_{f,t}]$.

The solution of this optimization problem cannot generally be found analytically, and in the case where both $\mathbf{H}$ and $\mathbf{W}$ are unknown, the problem is non-convex. A local optimum can be found iteratively by using a majorization-minimization (MM) method [17] that we refer to as KL-NMF.

The typical pipeline to perform speaker-dependent semi-supervised NMF-based speech enhancement in the presence of two sources, say speech and noise, follows the process detailed in [14], where the method is derived from the equivalent perspective of PLCA:

1. Compute the spectrogram $\mathbf{S}_S$ from the speech training data.
2. Factorize the spectrogram $\mathbf{S}_S \approx \mathbf{W}_S\tilde{\mathbf{H}}_S$ by minimizing $D(\mathbf{S}_S||\mathbf{W}_S\tilde{\mathbf{H}}_S)$ using KL-NMF and store the speech model $\mathbf{W}_S$.
3. Compute the spectrogram $\mathbf{S}$ of the test mixture signal.
4. Learn concurrently the speech and noise activations $\mathbf{H}_S$ and $\mathbf{H}_N$ as well as the noise model $\mathbf{W}_N$ from the mixture spectrogram $\mathbf{S}$ while keeping $\mathbf{W}_S$ fixed by minimizing $D(\mathbf{S}||\mathbf{W}_S\mathbf{H}_S + \mathbf{W}_N\mathbf{H}_N)$ using KL-NMF.
5. Construct estimated spectrograms for each source $\hat{\mathbf{S}}_i = \mathbf{W}_i\mathbf{H}_i$ for $i = S, N$.
6. Construct a time-frequency masks from the $\hat{\mathbf{S}}_i$ and extract the estimated STFTs of each source through Wiener filtering of the mixture STFT $\mathbf{X}$:

$$\hat{\mathbf{X}}_S = \frac{\hat{\mathbf{S}}_S}{\hat{\mathbf{S}}_S + \hat{\mathbf{S}}_N}\mathbf{X} \qquad \hat{\mathbf{X}}_N = \frac{\hat{\mathbf{S}}_N}{\hat{\mathbf{S}}_S + \hat{\mathbf{S}}_N}\mathbf{X}$$

7. Compute the inverse STFT of $\hat{\mathbf{X}}_i$ $(i = S, N)$ to get an estimate of each source.

Algorithm 1 Online Block KL-NMF

inputs $\mathbf{s}_t$, $[\mathbf{s}_1 \dots \mathbf{s}_L]$, $\mathbf{W}_i$, $[\mathbf{h}_{i,1} \dots \mathbf{h}_{i,L}]$ for $i = S, N$
set $\mathbf{S} = \left[\mathbf{s}_1 \dots \mathbf{s}_L\ \mathbf{s}_t\right]$
initialize $\mathbf{h}_{i,t}$ randomly
set $\mathbf{H}_i = \left[\mathbf{h}_{i,1} \dots \mathbf{h}_{i,L}\ \mathbf{h}_{i,t}\right]$ and $\mathbf{H} = \begin{bmatrix}\mathbf{H}_S \\ \mathbf{H}_N\end{bmatrix}$
set $\mathbf{W} = \left[\mathbf{W}_S\ \mathbf{W}_N\right]$ (assuming $1^T\mathbf{W} = 1$)
repeat
  $\mathbf{V} \leftarrow \mathbf{S}\ ./\ (\mathbf{WH})$
  $\widetilde{\mathbf{W}}_N \leftarrow \mathbf{W}_N\ .*\ (\mathbf{V}\mathbf{H}_N^T)$
  $\mathbf{h}_t \leftarrow \mathbf{h}_t\ .*\ (\mathbf{W}^T\mathbf{v}_t)$
  for $g = 1 : G$ **do**
    $\mathbf{h}_{S,t}^{(g)} \leftarrow \mathbf{h}_{S,t}^{(g)}\ ./\ \left(1 + \lambda/(\epsilon + ||\mathbf{H}_S^{(g)}||_1)\right)$
  end for
  $\mathbf{W}_N \leftarrow \widetilde{\mathbf{W}}_N\ ./\ (11^T\widetilde{\mathbf{W}}_N)$ (renormalize $\mathbf{W}_N$)
until convergence
return $\mathbf{W}_N$, $\mathbf{h}_{i,t}$ and $[\mathbf{h}_{i,1} \dots \mathbf{h}_{i,L}]$ for $i = S, N$

$.*$ and $./$ denote component-wise multiplication and division.

This method relies on the availability of isolated training data for the speaker in the mixture. A similar method can be derived for the case in which isolated training data is available for the noise part by swapping the two sources. Since those frameworks are then dependent on either the specific speaker or noise presented in the training examples, they can have trouble generalizing to unseen speakers and noise types.

In the USM framework [13], we leverage the fact that we expect some degree of similarity between groups of speakers allowing the spectral features of one speaker (or a few speakers) to explain the spectrograms of similar unseen speakers. Then, we change steps 1. and 2. by learning the speech models $\mathbf{W}_S^{(g)}$ $(g = 1 \dots G)$ of G distinct speakers and store them together as a USM $\mathbf{W}_S = [\mathbf{W}_S^{(1)} \dots \mathbf{W}_S^{(G)}]$. Under our hypothesis, we would expect that only a few speaker models (the most similar) need to be active in order to explain the spectrogram of an unseen speaker. We then learn the noise model $\mathbf{W}_N$ and the activations $\mathbf{H}$ by solving the following optimization problem:

$$\underset{\mathbf{W}_N,\mathbf{H}\geq 0}{\text{minimize}}\ D(\mathbf{S}||\mathbf{WH}) + \lambda\sum_{g=1}^{G}\log(\epsilon + ||\mathbf{H}_S^{(g)}||_1) \tag{4}$$

The second term is a regularization term given as a function of the activations $\mathbf{H}_S^{(g)}$ of the different speech models for $g = 1 \dots G$ (with an additive term ϵ in order to avoid $\log(0)$ cases). Its function is to enforce block sparsity, meaning that it encourages most of the speech models to have their activations $\mathbf{H}_S^{(g)}$ equal to zero in order to match our intuition that only a few speakers need to be active to explain the speech signal. The sparsity parameter λ controls the weight of the sparsity condition. High sparsity parameters encourage the selection of fewer models, resulting in better separation, but sometimes at the cost of higher artifacts. We refer to the method solving this optimization problem as Block KL-NMF [13].

2.2. Proposed method

In the context of an online method, we want to process the spectrogram one frame at a time in order to minimize the delay between input and output. Following an approach similar to the on-

line method presented in [8], we gather a buffer of L previously processed spectrogram frames, for which a noise model and the different activations have been estimated at an earlier iteration. The idea is then to factorize $\mathbf{S} = [\mathbf{s}_1 \dots \mathbf{s}_L\ \mathbf{s}_t]$, concatenation of the L previously processed frames $\mathbf{s}_1 \dots \mathbf{s}_L$ with the current frame $\mathbf{s}_t$ at time t. With $\mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_L\ \mathbf{h}_t]$, we keep fixed the activations of the buffer frames $\mathbf{h}_1, \dots, \mathbf{h}_L$ and the speech model (the USM $\mathbf{W}_S = [\mathbf{W}_S^{(1)} \dots \mathbf{W}_S^{(G)}]$), and update only the activations of the current frame $\mathbf{h}_t$ and the noise model $\mathbf{W}_N$. The optimization problem we solve for each new frame is then given by:

$$\underset{\mathbf{W}_N, \mathbf{h}_t \geq 0}{\text{minimize}}\ D(\mathbf{S}||\mathbf{W}\mathbf{H}) + \lambda \sum_{g=1}^{G} \log(\epsilon + ||\mathbf{H}_S^{(g)}||_1) \qquad (5)$$

Starting from the Block KL-NMF iterative method associated with USM offline separation [13], we derive the MM iterative method corresponding to our new optimization problem as presented in Algorithm 1. We refer to this method as online Block KL-NMF.

The outline of the method to perform the online separation using the USM is as follows:

- Perform the offline training of the USM (prior to deployment):
  1. Compute the spectrograms $\mathbf{S}_S^{(g)}$ from the speech data of each speaker g in the training set.
  2. Factorize the spectrograms $\mathbf{S}_S^{(g)} \approx \mathbf{W}_S^{(g)}\tilde{\mathbf{H}}_S^{(g)}$ for each distinct speaker $g = 1 \dots G$ using KL-NMF and store the learned speech models $\mathbf{W}_S^{(g)}$ as USM $\mathbf{W}_S = [\mathbf{W}_S^{(1)} \dots \mathbf{W}_S^{(G)}]$.
- Once deployed, we perform online speech enhancement. For each time frame of the mixture signal:
  1. Compute the corresponding spectrogram frame $\mathbf{s}_t$.
  2. Factorize $\mathbf{S} = [\mathbf{s}_1 \dots \mathbf{s}_L\ \mathbf{s}_t]$ by solving the optimization in Equation (5) using online Block KL-NMF.
  3. Recover the estimated speech $\hat{\mathbf{s}}_{S,t} = \mathbf{W}_S\mathbf{h}_{S,t}$ and noise frame $\hat{\mathbf{s}}_{N,t} = \mathbf{W}_N\mathbf{h}_{N,t}$ (from current $\mathbf{W}_N$).
  4. Construct spectral masks and extract the estimated STFT frame of each source through Wiener filtering of the mixture STFT frame $\mathbf{x}_t$:
  $$\hat{\mathbf{x}}_{S,t} = \frac{\hat{\mathbf{s}}_{S,t}}{\hat{\mathbf{s}}_{S,t} + \hat{\mathbf{s}}_{N,t}}\mathbf{x}_t \qquad \hat{\mathbf{x}}_{N,t} = \frac{\hat{\mathbf{s}}_{N,t}}{\hat{\mathbf{s}}_{S,t} + \hat{\mathbf{s}}_{N,t}}\mathbf{x}_t$$
  5. Compute the inverse STFT using the new frame $\hat{\mathbf{x}}_{i,t}$ $(i = S, N)$ and use overlap-and-add to obtain an new estimated segment for each source.
  6. Update the buffer containing the previously processed frames (adding the new frame and dropping the oldest one). Store the noise model $\mathbf{W}_N$ and the activations of the buffer frames to process the next frame.

2.3. Initialization

As the USM method is strongly non-convex due to the block sparsity condition, we need to ensure that the method is not trapped in a bad local minimum when filling the first buffer. To do so, we initialize the method by first storing an entire buffer of signal. We then perform offline separation on that segment by running 10 MM iterations of Block KL-NMF using the same parameter settings as the proposed online method. The estimated noise model and activations are then used as initial values for the next processed frame in the proposed method.

3. EXPERIMENTS

In this section, we determine optimal parameter settings for our method and compare its performance with offline NMF-based methods, the equivalent speaker-dependent NMF-based method, and several existing methods from speech enhancement literature.

3.1. Data and performance metrics

Our speech data is comprised of data from TIMIT dataset [18]. Our noise dataset is comprised of the following three existing noise datasets, for a total of 48 noise environments: the NOISEX-92 database [19], which contains primarily stationary noises, the dataset used in [8], which contains several examples of highly non-stationary noises, and the DEMAND database [20], which contains recordings of real-world environments.

We train a USM (as explained in Section 2) using data from 50 speakers such that each model in the USM is learned from the data (approximately 30 seconds) of a distinct speaker. A development set and a test set are built from data corresponding to a different set of 192 speakers (half male/half female), each speaker is mixed with one of the 48 noise examples to form 192 noisy mixtures with a signal-to-noise ratio of 0dB (speech and noise at the same energy level). As a result, each noise type is associated with exactly 4 mixtures (2 male/2 female). The development set consists of the mixtures associated with a third of the noise examples (16 noises/64 mixtures), and the test set of the remaining mixtures (32 noises/128 mixtures). By construction, the training, development, and test sets do not share any speaker or noise types. The mixtures are 12 seconds long, using concatenated segments from TIMIT as speech. Our files have a sampling rate of 16kHz, and the spectrograms are computed using a Hamming window of length 64ms (1024 samples) with a step size of 32ms (512 samples) and a zero-padding factor of 2. We run the online method with a buffer of 60 frames, corresponding to about 2 seconds of signal. Many existing techniques require such an initial buffer to contain isolated noise. Our method has no such requirement. The results are simply suboptimal in this region due to boundary effects. For each speaker, the TIMIT data that was not used to generate the mixture is used as isolated speaker-dependent training data for the speaker-dependent NMF-based baselines.

To evaluate the quality of the method, we use 4 standard metrics. The Signal-to-Noise Ratio (SNR), Segmental SNR (SegSNR) and Perceptual Evaluation of Speech Quality (PESQ) are three standard measures of speech enhancement quality. PESQ scores are mapped to a Mean Opinion Score (MOS) scale between 1 and 5 (higher is better). We use the implementations of those metrics given in [1]. The Source-to-Distortion Ratio (SDR) corresponds to the overall score defined in the BSS evaluation metrics [21]. It evaluates the quality of source separation methods, aggregating the distortions introduced by the interfering source and the audio artifacts generated by the method. We report here the SDR associated with the speech track in order to evaluate its quality.

3.2. Parameter Determination

To determine the best set of parameters for our method, we run it on the development set for the following parameter combinations: $N_S = 6, 10, 20, 40$; $K_S = 5, 10, 20, 40$; $K_N = 5, 10, 20, 40$; $\lambda = 8, 16, 32, 64$. Additionally, we determine the optimal number of MM iterations using the procedure described in [22], choosing the number of iterations that leads to the best average SDR score for speech in the development set. We run our method for $5, 10, 15, 20$

	Online	Offline
Universal models	$N_S = 20$ $K_S = 10$ $K_N = 10$ $\lambda = 64$ MM iter./frame $= 20$	$N_S = 40$ $K_S = 20$ $K_N = 150$ $\lambda = 256$ MM iter. $= 20$
Speaker dependent	$K_S = 10$ $K_N = 5$ MM iter./frame $= 5$	$K_S = 20$ $K_N = 150$ MM iter. $= 10$

Table 1. Parameters for the NMF-based methods

MM iterations for each new frame. Although the optimal group sparsity parameter λ ideally should depend on the SNR of the mixture, and the number of noise spectral features K_N on the complexity of the noise structure, we determine a single optimal value over all the examples in the interest of automating the system.

3.3. Baselines

We compare the proposed method to REPET-SIM [11] as another method performing speaker and noise independent online speech enhancement. Based on the assumptions mentioned in Section 1, this method relies on the building of a similarity matrix between the frames of the spectrogram to identify similar frames in the spectrogram followed by median filtering to average those frames and obtain the underlying repetitive pattern. Those patterns are then used to reconstruct the background noise signal through time-frequency masking. The enhanced speech estimate is finally obtained as the residual between the mixture signal and the noise estimate. The parameters used for this method are taken from [11]. We use the same STFT parameters that we used for the proposed method.

We also compare our results to several traditional speech enhancement methods, namely multi-band spectral subtraction [2], Wiener filtering [3], logarithmic Minimum Mean Square Error (log-MMSE) [4] and the KLT subspace method [5]. We use directly the implementations provided in [1]. All these methods are online methods that require adding a segment of signal without speech at the beginning of each mixture to learn a model for the noise environment, which is assumed to be quasi-stationary.

Finally, we compare the results of the proposed method with three other NMF-based methods: speaker and noise independent offline [13], speaker-dependent online [8], and speaker-dependent offline speech enhancement [14]. As for the proposed method, the noise model is learned from the mixture spectrogram. In the offline methods, the spectrogram of the entire signal is factorized at once. For the speaker-dependent methods, the speech model is learned using isolated training data of the given speaker as opposed to using the USM. These three methods are expected to perform better than the proposed method. Our evaluation aims at demonstrating that the decrease in speech enhancement performance is minimized compared to those methods. We perform parameter sweeps on the development set for each of these methods to find optimal parametrization and allow for a fair comparison.

3.4. Experimental results

The parameters used for the optimal result for each NMF-based method are given in Table 1. The average metrics of the different methods on the test set are presented in Table 2. We see that with respect to all metrics, the proposed method performs comparably to the online method for which speaker-dependent training data was available. A reason for this could be that while each individual speech model of the USM is an imperfect approximation, the combination of those models can allow for a more flexible model by using more than one model to approximate an unseen speaker. The performance difference between the offline and online methods could be explained by the availability of a larger amount of information (the entire spectrogram) to learn the noise structure in the offline case.

Method	SDR (dB)	PESQ (MOS)	SNR (dB)	SegSNR (dB)
Offline Speaker-dependent	10.96	1.72	10.55	4.37
Offline Universal models	10.9	1.81	10.02	4.1
Online Speaker-dependent	8.16	1.51	8.08	2.85
Online (Proposed) Universal models	8.26	1.53	8.19	2.87
REPET-SIM [11]	6.95	1.41	7.29	1.75
Spectral Subtraction [2]	3.27	1.34	1.59	-1.78
Wiener filtering [3]	3.87	1.33	3.51	0.75
log-MMSE [4]	5.52	1.46	5.53	1.78
KLT [5]	3.92	1.25	3.35	0.77

Table 2. Average evaluation metrics from the test data results (For all metrics: higher is better).

Our method also significantly outperforms traditional speech enhancement methods with respect to all metrics, in part because the assumption of a stationary noise environment limits the ability of those methods to handle non-stationary noises. In addition, our method outperforms REPET-SIM by a small margin with respect to all metrics as well (+1.3dB SDR, +0.12 MOS PESQ, +0.9dB SNR, +1.1dB SegSNR). Upon inspection of our data, we observe that our method could much better approximate highly non-stationary noises than REPET-SIM, as those noises generally do not exhibit the repetitive structure expected by this method. These results emphasize the advantage of having an explicit modeling of the speech structure even in the absence of speaker-dependent training data in order to avoid assumptions on the noise structure.

4. CONCLUSION

In this paper, we presented a speaker and noise independent online speech enhancement method extending USM-based speech enhancement to the online setting. This framework allowed us to match the speech patterns of unseen speakers without the need for speaker-dependent training data. This method was tested on mixtures derived from standard datasets, and it demonstrated similar performance to the speaker-dependent implementation of the online method.

Potential variants of this method include the development of an online voice activity detection method by adapting the offline method presented in [15]. The addition of temporal modeling in other NMF-based methods [23, 24] has been shown to improve noticeably source separation quality. Other potential improvements include the adaptive selection of some of the parameters, such as the number of noise spectral features K_N or the sparsity parameter λ. Future developments of the presented method could take advantage of the combination of USMs with such extensions.

5. REFERENCES

[1] P. C. Loizou, *Speech enhancement : theory and practice*, vol. 30, CRC Press, Boca Raton, FL, 2007.

[2] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2002, vol. 4, pp. 4164–4164.

[3] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1996, vol. 2, pp. 629–632.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[5] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, July 2003.

[6] A. Ozerov and E. Vincent, "Using the FASST source separation toolbox for noise robust speech recognition," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, Florence, Italy, September 2011, pp. 86–87.

[7] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation*, Berlin, Heidelberg, 2012, LVA/ICA'12, pp. 322–329, Springer-Verlag.

[8] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments.," in *Proceedings of INTERSPEECH 2012*. 2012, ISCA.

[9] Ö. Ylmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830–47, 2004.

[10] F. Nesta and M. Matassoni, "Robust automatic speech recognition through on-line semi blind source extraction," in *CHIME 2011 Workshop on Machine Listening in Multisource Environments*, Florence, Italy, September 2011, pp. 18–23.

[11] Z. Rafii and B. Pardo, "Online repet-sim for real-time speech enhancement," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 848–852.

[12] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix.," in *Proceedings of the 2012 ISMIR Conference*, 2012, pp. 583–588.

[13] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[14] P. Smaragdis, B. Raj, and M. V. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation*, Berlin, Heidelberg, 2007, ICA'07, pp. 414–421, Springer-Verlag.

[15] F. G. Germain, D. L. Sun, and G. J. Mysore, "Speaker and noise independent voice activity detection.," in *Proceedings of INTERSPEECH 2013*. 2013, pp. 732–736, ISCA.

[16] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[17] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of Neural Information Processing Systems (NIPS)*, T.K. Leen, T.G. Dietterich, and V. Tresp, Eds. 2000, vol. 13, pp. 556–562, MIT Press.

[18] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, *DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM*, National Institute of Standards and Technology, NISTIR 4930, 1993.

[19] A. Varga and H. 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, 1993.

[20] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," 2013.

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

[22] F. G. Germain and G. J. Mysore, "Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1284–1288, 2014.

[23] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *Proceedings of the 9th international conference on Latent variable analysis and signal separation*, Berlin, Heidelberg, 2010, LVA/ICA'10, pp. 140–148, Springer-Verlag.

[24] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 873–877.