2011 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) December 7-9, 2011

A New Approach for Speech Enhancement Based on a Constrained Nonnegative Matrix Factorization

Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon
Sound and Image Processing Lab.
KTH Royal Institute of Technology, Stockholm, Sweden
Email: [nmoh,gerkmann,leijon]@kth.se

***Abstract*—In this paper, a new approach is presented for single-channel speech enhancement which is based on Nonnegative Matrix Factorization (NMF). The proposed scheme combines the noise Power Spectral Density (PSD) estimation based on a constrained NMF and Wiener filtering to enhance the noisy speech. The imposed constraint is motivated by the time correlation of the underlying observations and enforces the NMF to give smoother estimates of the nonnegative factors. Compared to the standard NMF approach and Wiener filtering based on a recently developed noise PSD estimator, Source to Distortion Ratio (*SDR*) is improved for the evaluated noise types for different input *SNR*s.**

I. Introduction

The quality and pleasantness of speech may significantly deteriorate in the presence of background noise. The problem of speech enhancement under the additive noise assumption has been widely studied in the past, and is still an active field of research.

In this paper, we consider a supervised learning approach based on Nonnegative Matrix Factorization (NMF) to enhance the noisy speech signal. NMF finds a locally optimal solution to solve the matrix equation $X \approx TV$ under the nonnegativeness constraint on T and V. For NMF based speech enhancement or audio source separation, X is the magnitude or power spectrogram of the observed signal, where spectra are stored column-wise in X. NMF is applied to factorize the spectrogram into a matrix consisting of basis matrix, T, and a NMF coefficients matrix, V, which represents the activity of each basis vector over time.

NMF has been widely used for blind source separation (BSS), e.g. [1], [2], [3]. In general, two approaches are used for BSS: in the first approach, after performing the factorization each separated source is obtained as a weighted sum of the basis vectors, where weighting factors are given by the NMF coefficients matrix V [1], [2]. In another approach, a soft mask is formed using the result of the factorization, and each separated source is obtained by a product of the mask and the observed matrix [3].

On the other hand, there are fewer studies that use NMF for speech denoising. In [4], [5], a constrained NMF is used to enhance the noisy speech; the constraint is motivated by the fact that human speech and some interference signals have different spectral structures, at least partially. Similar to the approach which is used in BSS, a weighted sum of the speech basis vectors are used to approximate the target speech in these papers. In [6], a semi-supervised algorithm is introduced which is based on nonnegative hidden Markov modeling [7], and a Wiener-type gain, similar to the one used in [3], is used to enhance the noisy speech. There is an important difference between BSS and speech enhancement: in BSS, we aim at separating all the sources as well as possible while in speech enhancement we are not interested in estimating the noise as a separate target source.

NMF based speech enhancement algorithm is usually based on a supervised learning; hence, the NMF based algorithm consists of two steps: first a set of basis vectors are found for the given training signals. Secondly, these basis vectors are used in the separation or enhancement step. Consequently, improvements can be made on either of these stages.

We will use a standard NMF in the training part, and will contribute to make a better factorization in the enhancement phase. The Wiener-type gain for the enhancement of the noisy speech is in fact a spectral subtraction algorithm in which a periodogram type noise PSD estimate is used. We propose an algorithm to derive a better estimate for the noise PSD. To do so, first a constrained NMF is used which enforces the factorization to give a smooth estimate of the noise signal; this constraint is motivated by the observation that noise is often more stationary than the speech signal. Secondly, to get the noise PSD we smooth the estimated noise periodogram across time to reduce its variance. The obtained noise PSD is then used in combination with the Wiener filtering to enhance the noisy speech. The performance of the proposed method is measured with different instrumental measures including Source to Distortion Ratio (*SDR*), segmental speech *SNR*, and segmental noise reduction.

II. Notation and Basic Concepts

To refer to the (k,τ)th entry of a matrix X, we use $X(k,\tau)$; $\mathbf{x}_{\cdot\tau}$ denotes the τth column of the matrix X, and $\mathbf{x}(k)$ denotes the kth element of the vector $\mathbf{x}$. Let $Y(k,\tau)$ denote the DFT coefficient for frequency bin k, and time-frame τ. Assuming that noisy speech consists of speech degraded by additive noise, we have $Y(k,\tau) = S(k,\tau) + N(k,\tau)$, where Y, S, N are the noisy speech, clean speech and the noise DFT coefficients, respectively. We use the magnitude-squared DFT coefficients ($|\cdot|^2$) as the observation in the NMF.

Given the nonnegative matrix X, there are different algorithms to perform the factorization [8], [9]. Here, we use

978-1-4577-2166-3/11/$26.00 ©2011 IEEE

generalized Kullback-Leibler divergence as the cost function:

$$D_{KL}(X\|TV) = \sum_{k,\tau}(X(k,\tau)\log\frac{X(k,\tau)}{\Lambda(k,\tau)}+\Lambda(k,\tau)-X(k,\tau)), \quad (1)$$

where $\Lambda = TV$. Factors T, V are found by iterating the following multiplicative rules [10]:

$$\begin{aligned} T(k,i) &\leftarrow T(k,i)\frac{\sum_\tau V(i,\tau)(X(k,\tau)/\Lambda(k,\tau))}{\sum_p V(i,p)}, \\ V(i,\tau) &\leftarrow V(i,\tau)\frac{\sum_k T(k,i)(X(k,\tau)/\Lambda(k,\tau))}{\sum_q T(q,i)}, \end{aligned} \quad (2)$$

after updating T, the columns of T are normalized such that each column sums to 1.

III. State-of-the-art Speech Enhancement Based on NMF (Stand-NMF)

NMF based speech enhancement algorithm consist of a training and an enhancement phase. For both steps, the given time-domain signal is segmented, windowed, and transformed into the frequency domain to obtain the spectrogram. During the training phase, NMF is applied to the power spectrogram of the clean speech and noise signals to obtain the speech basis matrix, T_S, and noise basis matrix, T_N:

$$(T_S, V) = \arg\min_{T,Z} D_{KL}\left(|S_{train}|^2 \|TZ\right), \quad (3)$$

$$(T_N, W) = \arg\min_{T,Z} D_{KL}\left(|N_{train}|^2 \|TZ\right), \quad (4)$$

where S_{train} and N_{train} are the DFT coefficients of the clean speech and noise signals, respectively, and $|\,.\,|^2$ is an element-wise operator. Now, the basis matrix for the observed noisy speech, T, is obtained as: $T = (T_S \;\; T_N)$. In the enhancement phase, an overlap-add framework is utilized to process each frame of the noisy speech $(\mathbf{y}_{.\tau})$ separately. Keeping the basis matrix fixed, NMF is performed to obtain the NMF coefficients vector $\mathbf{u}_{.\tau}$:

$$\mathbf{u}_{.\tau} = \arg\min_{\mathbf{z}} D_{KL}\left(|\mathbf{y}_{.\tau}|^2 \|T\mathbf{z}\right). \quad (5)$$

The basic idea in (5) is to find a linear combination of the speech basis matrix, T_S, and noise basis matrix, T_N, which best approximates the noisy input $|\mathbf{y}_{.\tau}|^2$, such that: $|\mathbf{y}_{.\tau}|^2 \approx T_S\mathbf{v}_{.\tau}+T_N\mathbf{w}_{.\tau} = T\mathbf{u}_{.\tau}$, where $\mathbf{u}_{.\tau} = \left(\mathbf{v}_{.\tau}^\top \; \mathbf{w}_{.\tau}^\top\right)^\top$ ($\top$ denotes the transpose).

An estimate of the speech magnitude-squared DFT coefficients is now obtained by multiplying a Wiener-type gain to the observation as:

$$\widehat{|S(k,\tau)|^2} = \frac{\lambda^s_{.\tau}(k)}{\lambda^s_{.\tau}(k)+\lambda^n_{.\tau}(k)} \times |Y(k,\tau)|^2, \quad (6)$$

where $\lambda^s_{.\tau} = T_s\mathbf{v}_{.\tau}$ and $\lambda^n_{.\tau} = T_N\mathbf{w}_{.\tau}$.

The time-domain enhanced speech can be obtained by inverse DFT transform using the noisy phase information and an overlap-add resynthesis. Eq. (6) is in fact an instantaneous power spectral subtraction (PSS) algorithm because:

$$\widehat{|N(k,\tau)|^2} = \frac{\lambda^n_{.\tau}(k)}{\lambda^s_{.\tau}(k)+\lambda^n_{.\tau}(k)} \times |Y(k,\tau)|^2, \quad (7)$$

$$\begin{aligned} \widehat{|S(k,\tau)|^2} &= (1 - \frac{\lambda^n_{.\tau}(k)}{\lambda^s_{.\tau}(k)+\lambda^n_{.\tau}(k)}) \times |Y(k,\tau)|^2 \\ &= |Y(k,\tau)|^2 - \widehat{|N(k,\tau)|^2}, \end{aligned} \quad (8)$$

thus, the standard NMF can be seen as performing PSS based on an estimate of the noise periodogram.

IV. Proposed Algorithm

In this section, we describe an approach to get an improved estimate of the noise PSD; then, the obtained noise PSD will be used in combination with the Wiener filter to enhance the noisy speech. The training phase of the proposed algorithm is identical to the one in Section III; therefore, we only consider the enhancement phase here.

A. Noise PSD Estimation

Keeping the same notations from Section III, and assuming some extent of stationarity of the noise, we can smooth the result of (7) across time to get a better noise PSD estimate as:

$$\widehat{\sigma^2_N}(k,\tau) = \beta\widehat{\sigma^2_N}(k,\tau-1) + (1-\beta)\widehat{|N(k,\tau)|^2}, \quad (9)$$

where $\widehat{|N(k,\tau)|^2}$ is given in (7), and $\widehat{\sigma^2_N}(k,\tau)$ denotes the estimated noise PSD for frequency bin k, and time-frame τ. Note that this kind of smoothing is a common technique to reduce the variance of the estimate in noise PSD estimation algorithms [11].

B. Constrained NMF (C-NMF)

Since human speech and some interference signals like babble noise have similar spectral structures, at least partially, the basis matrices of speech and these noise signals may be rather similar; as a result, the separation of the noisy speech into speech and noise components is difficult. Thus, there should be more constraints on the factorization to make it more robust. Here, we introduce a constraint which enforces the smoothness on the estimated noise signal. This is motivated by the fact that most of the noise signals are more stationary than the speech signal; hence, the consecutive estimates of the noise should be highly correlated. To do so, the following constrained NMF problem is considered to obtain the NMF coefficients vector $\mathbf{u}_{.\tau} = \left(\mathbf{v}_{.\tau}^\top \; \mathbf{w}_{.\tau}^\top\right)^\top$:

$$\mathbf{u}_{.\tau} = \arg\min_{\mathbf{z}} D_{KL}\left(|\mathbf{y}_{.\tau}|^2 \|T\mathbf{z}\right) + \alpha J(\mathbf{z}), \quad (10)$$

where $J(\mathbf{z})$ is a convex function of $\mathbf{z}$. In the optimization problem given in (10), the basis matrix T is constant and any scaling of the observation will be carried to the NMF coefficients vector $\mathbf{u}_{.\tau}$ (i.e. $a|\mathbf{y}_{.\tau}|^2 \approx T(a\mathbf{u}_{.\tau})$). Since the Kullback-Leibler divergence is scale variant [1] ($D_{KL}(ay\|ax) = aD_{KL}(y\|x)$), we should define a penalty

term $J(\mathbf{z})$ which also scales like the Kullback-Leibler divergence. This is important because the optimal value of α will be obtained using a cross-validation approach and the validation and test materials might have different scales, and the optimality of α should not be changed because of the difference in the scaling. Based on this explanation, we use a scale variant $J(\mathbf{z})$ in (10) for which: $J(a\mathbf{z}) \approx aJ(\mathbf{z})$. We also note that given the basis matrix T, the Kullback-Leibler divergence is a convex function of $\mathbf{z}$ and to preserve the convexity of the cost function, $J(\mathbf{z})$ is required to be a convex function. In the rest of this Section, we first introduce two alternative choices for $J(\mathbf{z})$, and then we explain the optimization procedure. In the following, we apply a constraint only on the part of the NMF coefficients vector which corresponds to the noise basis vectors.

1) Normalized Sum of the Squared Differences : The NMF coefficients corresponding to the noise signal exhibit smoothed variations during time according to the stationarity of the noise. Splitting $\mathbf{z}$ as: $\mathbf{z} = \left(\mathbf{z}_s^\top \ \mathbf{z}_n^\top\right)^\top$, and assuming that NMF coefficients corresponding to the noise signal ($\mathbf{z}_n$) have multivariate Gaussian distribution with diagonal covariance matrix and similar diagonal elements, the negative log likelihood of $\mathbf{z}_n$ is related to the sum of the squared differences between the elements of $\mathbf{z}_n$ and their mean values. Hence, we define the following normalized penalty term:

$$J(\mathbf{z}) = \frac{1}{\sigma_{\overline{\mathbf{w}}_{\cdot\tau}}} \sum_j \left(\mathbf{z}_\mathrm{n}(j) - \overline{W}(j,\tau)\right)^2, \qquad (11)$$

where $\overline{W}(j,\tau)$ is the mean activity level of the jth basis vector of the noise at time frame τ, and $\sigma_{\overline{\mathbf{w}}_{\cdot\tau}}$ is the rms value of $\overline{\mathbf{w}}_{\cdot\tau}$; these parameters are described in the following.

$\overline{W}(j,\tau)$ is estimated by smoothing the previous values of W across time as:

$$\overline{W}(j,\tau) = \gamma\overline{W}(j,\tau-1) + (1-\gamma)W(j,\tau-1). \qquad (12)$$

$\sigma_{\overline{\mathbf{w}}_{\cdot\tau}}$ is defined as:

$$\sigma_{\overline{\mathbf{w}}_{\cdot\tau}} = \sqrt{\frac{1}{I}\sum_{j=1}^{I}\overline{W}(j,\tau)^2},$$

where I is the number of the noise basis vectors (number of columns of T_N). By scaling $\mathbf{w}$, its mean value will be also scaled, and it is easy to verify that $J(a\mathbf{z}) = aJ(\mathbf{z})$.

In [2], a different form of sum of the squared errors was used as the penalty term for NMF based monaural sound source separation. Eq. (11) differs from the one introduced in [2] for two reasons:

1) Our penalty term penalizes the variations of noise NMF coefficients from their mean values while in [2] the differences between the consecutive NMF coefficients were penalized. Since the interfering signal in speech enhancement application is usually more stationary than the speech, using the mean values instead of the last neighboring coefficients statistically makes more sense and gives better results.
2) Because of the applied normalization, the introduced penalty term in [2] is scale invariant while (11) is scale variant; it means that the value of the penalty term is adjusted depending on the level of the noisy speech. Hence, after selecting an optimal value for α (which controls the effect of the penalty term) at one particular input level, regardless of the level of the input (e.g. regardless of the level of the noise in the noisy speech) the penalty term will be scaled as the Kullback-Leibler divergence and its influence will remain optimal.

We will also need the derivative of $J(\mathbf{z})$ in the optimization which is calculated as:

$$\phi_\mathbf{z} = \frac{\partial J(\mathbf{z})}{\partial \mathbf{z}} = \left(0^\top \quad \frac{2}{\sigma_{\overline{\mathbf{w}}_{\cdot\tau}}}(\mathbf{z}_\mathrm{n} - \overline{\mathbf{w}}_{\cdot\tau})^\top\right)^\top, \qquad (13)$$

where 0 is a zero vector that has the same number of elements as the number of the speech basis vectors (number of columns of T_S).

2) Kullback-Leibler Divergence: As an alternative, we use the Kullback-Leibler divergence (1) between the NMF coefficients vector and its mean value as:

$$J(\mathbf{z}) = D_{KL}(\mathbf{z}_\mathrm{n} \| \overline{\mathbf{w}}_{\cdot\tau}). \qquad (14)$$

The scaling problem is solved using divergence with no more efforts. $\overline{\mathbf{w}}_{\cdot\tau}$ is computed as (12). The derivative of $J(\mathbf{z})$ is given as:

$$\phi_\mathbf{z} = \frac{\partial J(\mathbf{z})}{\partial \mathbf{z}} = \left(0^\top \quad \left(\log\frac{\mathbf{z}_\mathrm{n}}{\overline{\mathbf{w}}_{\cdot\tau}}\right)^\top\right)^\top, \qquad (15)$$

3) Optimization Procedure: The optimization problem in (10) is solved using an iterative approach. The update rule for this iterative optimization is given as [8]:

$$\mathbf{z}(i) \leftarrow \mathbf{z}(i)\frac{\sum_k T(k,i)\left(|Y(k,\tau)|^2/\lambda(k)\right)}{\max\left(\sum_q T(q,i) + \alpha\phi_\mathbf{z}(i), \epsilon\right)}, \qquad (16)$$

where $\lambda = T\mathbf{z}$, and ϵ is a small positive number. $\phi_\mathbf{z}$ is defined as (13) or (15). The value of $\mathbf{z}$ is assigned to $\mathbf{u}_{\cdot\tau}$ after 150 iterations. After obtaining $\mathbf{u}_{\cdot\tau}$, the estimated noise magnitude-squared DFT coefficients are obtained using (7), and the noise PSD is estimated according to (9). It should be noted that the extra complexity of the proposed scheme, which is due to the calculation of $\phi_\mathbf{z}$, is ignorable compared to the original cost of the iterative minimization of (5).

V. Evaluation

The proposed algorithm was used to estimate the noise PSD which was used in combination with a Wiener filter to enhance the noisy speech; In our simulation, using the first penalty term (11) resulted to slightly better results than the second one (14), and those results are reported here as the proposed algorithm. In addition, noise PSD was estimated using a MMSE based approach [12] which is one of the best algorithms for this purpose [11], and the same Wiener filter was used for the enhancement; this approach is called

Wiener-MMSE in the following. The Wiener filter was implemented using the decision-directed approach [13] with the same parameters $10\log_{10}(\xi_{min}) = -25$dB, $\alpha = 0.98$ and $20\log_{10}(G_{min}) = -20$ dB for both approaches; ξ_{min} is the lower bound for the a priori SNR, and G_{min} is the lower bound for the Wiener gain. Furthermore, the standard NMF approach from Section III (Stand-NMF) is also included in the evaluation; In performing (6), applying gain limit like the one applied to the Wiener gain for the other two approaches, degraded the performance and hence in the evaluations no gain limit was used for this approach.

We used speech from the Grid Corpus and noise from the NOISEX-92 databases. All the signals are down-sampled to 16 KHz. The speech is degraded by adding babble noise or factory noise at 3 different *SNR*s: 0 dB, 5 dB, and 10 dB. A separate model is trained for each noise type; if noise type in not known a priori, some adaptive or semi-supervised approaches like [6] have to be used. One *speaker independent* model is trained for the speech signal; this model was trained on a mixed group of male and female speakers, none of which were in the test set. Speech signals from 24 speakers (12 speaker per gender), and 8 sentences from each speaker were concatenated to obtain the training data for this model. For all the approaches a similar test set was used, which consists of sentences from 4 male and 4 female speakers, none of which were in the train set, and 10 sentences for each speaker. The part of the noise signal which was used for the test purposes was not used in the training. The results are averaged over all the test set.

For the speech model and noise model, 60 and 100 basis vectors are trained, respectively. The following parameters are used in the simulations: $\alpha = 0.004$ in (10) which is chosen using a cross-validation technique, $\gamma = \beta = 0.95$ in (9, 12). The time frames have a length of 512 samples with 50% overlap, and are windowed using a Hann window.

The performance of the speech enhancement algorithms are evaluated using the Source to Distortion Ratio *(SDR)* which is defined as:

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{artifact}\|^2},$$

where s_{target}, e_{interf} and $e_{artifact}$ are target time-domain speech signal, interference, and artifact error terms defined in [14], and $\|\cdot\|^2$ denotes the energy. *SDR* represents the overall quality of speech when reducing noise and absence of artifacts are equally important. In order to analyze the results more specifically, the segmental speech *SNR*(SNR_{sp}) , and the segmental noise reduction (*SegNR*) are also measured as [15]:

$$SNR_{sp} = \frac{1}{T}\sum_{\tau=1}^{T} 10\log_{10}\left(\frac{\sum_{i=1}^{I} s_{i+\tau I}^2}{\sum_{i=1}^{I}(s_{i+\tau I} - \tilde{s}_{i+\tau I})^2}\right),$$

$$SegNR = \frac{1}{T}\sum_{\tau=1}^{T} 10\log_{10}\left(\frac{\sum_{i=1}^{I} n_{i+\tau I}^2}{\sum_{i=1}^{I} \tilde{n}_{i+\tau I}^2}\right),$$

where I denotes the length of the frame, and T the number of frames; these measures are obtained in a shadow filtering framework: the Wiener filter is computed from the noisy speech signal $(s+n)$ and is used to obtain $\tilde{s}, \tilde{n}$. $\tilde{s}$ is the output of the enhancement system when the clean speech s, is the input to the filter; similarly, $\tilde{n}$ is the output of the enhancement system when only the noise n, is the input to the filter.

Input *SNR*	Stand-NMF	Proposed	Wiener-MMSE
0 dB	0.2	**2.6**	1.5
5 dB	-0.25	**2.3**	1.6
10 dB	-1.7	1	**1.1**

(a) Babble noise

Input *SNR*	Stand-NMF	Proposed	Wiener-MMSE
0 dB	1.7	**3.8**	2.1
5 dB	1.1	**3.2**	2.2
10 dB	-0.4	**1.6**	1.4

(b) Factory noise

TABLE I: SDR improvements in dB for babble and factory noises

The noise tracking performance is evaluated using an averaged log distance $LogErr_{mean}$ defined as:

$$LogErr_{mean} = \frac{1}{KT}\sum_{k=1}^{K}\sum_{\tau=1}^{T}\left|10\log_{10}\left[\frac{\left[\sigma_N^2\right]_{k,\tau}}{\left[\widehat{\sigma_N^2}\right]_{k,\tau}}\right]\right|,$$

where σ_N^2 is the reference noise PSD, and is obtained by smoothing the noise periodograms across time:

$$\left[\sigma_N^2\right]_{k,\tau} = 0.95\left[\sigma_N^2\right]_{k,\tau-1} + 0.05\left|N_{k,\tau}\right|^2. \quad (17)$$

A. Performance Evaluation

Table I shows the *SDR* improvement for different algorithms. Different rows in Table I present the results for different input *SNR*s. The proposed algorithm results in the best scores for most of the cases, especially its performance excels significantly at low input *SNR*s. The Stand-NMF approach consistently gives worse results than the proposed and the Wiener-MMSE approaches. It has to be emphasized that NMF-based approaches are based on a *speaker independent* model, and by using a *speaker dependent* model the results of the both NMF-based approaches would be increased; however, to have a fair comparison between all the algorithms, we only consider the results of the *speaker independent* model here.

More insights can be gained by looking at Table II which shows the Segmental Noise Reduction, *SegNR*, and Speech *SNR*, SNR_{sp}, for babble noise at different input *SNR*s. For both measures a high value is desired, and SNR_{sp} is inversely proportional to the speech distortion. Wiener-MMSE approach provides a high *SegNR* while SNR_{sp} remains quite small especially at low input *SNR*s; On the other hand, Stand-NMF results to high SNR_{sp} and low *SegNR*. The proposed algorithm makes a compromise between two measures and results to higher quality compared to the other approaches. Informal listening tests verified these results.

Finally, Table III shows some detailed results for the performance of the noise tracking part. A small value for

Input *SNR*	Measure	Stand-NMF	Proposed	Wiener-MMSE
0 dB	SNR_{sp}	8.9	9.1	5.4
	SegNR	4.3	6.7	10.5
5 dB	SNR_{sp}	9.5	11.4	8.8
	SegNR	4.4	6.4	8.8
10 dB	SNR_{sp}	10.1	13.4	13.1
	SegNR	4.4	6.3	6.7

TABLE II: Segmental speech *SNR* (SNR_{sp}) and Segmental Noise Reduction (*SegNR*) in dB for babble noise in different input *SNR*s

Input *SNR*	Proposed	Wiener-MMSE
0 dB	**2.3**	2.6
5 dB	2.8	**2.4**
10 dB	4.1	**2.3**

TABLE III: $LogErr_{mean}$ in dB for babble noise

$LogErr_{mean}$ is desired. Stand-NMF is not shown in this table because this approach provides an estimate of the noise magnitude-squared DFT coefficients and comparing it to the (smoothed) noise PSD reference is not fair (it results in large error values). The proposed approach has smaller error for 0 dB *SNR* while MMSE has lower errors for higher input *SNR*s. This comparison shows that the estimated noise PSD is less similar to the reference noise PSD compared to the estimate given by MMSE approach [12] particularly at 10 dB input *SNR*. However, this had a small effect in the performance of the enhancement procedure since both approaches resulted to quite similar outputs, having around 1 dB *SDR* improvement.

VI. Conclusions

An approach was proposed for estimating the noise PSD based on a constrained NMF. The applied constraint penalized the variations of the NMF coefficients vectors from their mean values, resulting to smoother estimates which in turn takes into account the time correlations of the noise signal. The estimated noise PSD was used in a Wiener filtering framework to enhance the noisy speech. The performance of the speech enhancement was evaluated using different instrumental measures including *SDR*, segmental noise reduction, and segmental speech *SNR*. The simulations show that the proposed approach outperforms the standard NMF approach and MMSE based Wiener filtering for different input *SNR*s and considered noise types. A *SDR* improvement in the order of 2.6 dB was obtained using the proposed approach for a noisy speech degraded by an additive babble noise at 0 dB input *SNR*. Informal listenings verified the excellence of the proposed algorithm.

VII. Acknowledgements

This work was supported by the EU Initial Training Network AUDIS (grant 2008-214699).

References

[1] 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.

[2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, March 2010.

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE Int. Conf. ICASSP*, 2008.

[5] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Interspeech*, 2008, pp. 411–414.

[6] 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 *IEEE Int. Conf. ICASSP*, 2011.

[7] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden Markov modeling of audio with application to source separation," in *international conference on Latent Variable Analysis and Signal Separation*, 2010.

[8] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *IEEE Int. Conf. ICASSP*, 2006.

[9] C. Févotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *EUSIPCO*, 2009.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000.

[11] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *IEEE Int. Conf. ICASSP*, 2011.

[12] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *IEEE Int. Conf. ICASSP*, 2010.

[13] Y. Ephraim and I. Cohen, *Recent advancements in speech enhancement.* in The Electrical Engineering Handbook, CRC Press, 2005.

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

[15] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, 2005.