A NEW LINEAR MMSE FILTER FOR SINGLE CHANNEL SPEECH ENHANCEMENT BASED ON NONNEGATIVE MATRIX FACTORIZATION

Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon

KTH Royal Institute of Technology, Sound and Image Processing Lab., Stockholm, Sweden
{nmoh, gerkmann, leijon}@kth.se

ABSTRACT

In this paper, a linear MMSE filter is derived for single-channel speech enhancement which is based on Nonnegative Matrix Factorization (NMF). Assuming an additive model for the noisy observation, an estimator is obtained by minimizing the mean square error between the clean speech and the estimated speech components in the frequency domain. In addition, the noise power spectral density (PSD) is estimated using NMF and the obtained noise PSD is used in a Wiener filtering framework to enhance the noisy speech. The results of the both algorithms are compared to the result of the same Wiener filtering framework in which the noise PSD is estimated using a recently developed MMSE-based method. NMF based approaches outperform the Wiener filter with the MMSE-based noise PSD tracker for different measures. Compared to the NMF-based Wiener filtering approach, Source to Distortion Ratio (*SDR*) is improved for the evaluated noise types for different input *SNR*s using the proposed linear MMSE filter.

Index Terms— Speech enhancement, nonnegative matrix factorization, Linear MMSE filter

1. INTRODUCTION

In this paper, we consider a supervised 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 NMF basis vectors, T, and the NMF coefficients matrix, V, which represents the activity of each basis vector over time. NMF has been widely used for blind source separation (BSS) and speech enhancement recently; after performing NMF on the mixed observed speech signal, in general, the separated or enhanced signal is obtained using one of the following approaches: 1) as a weighted sum of the basis vectors, where weighting factors are given by the NMF coefficients matrix, V [1, 2, 3, 4]. 2) by a product of a Wiener-type soft mask and the observation matrix X [5]. In this paper we aim to obtain an optimal soft mask for enhancing the noisy signal.

We use a supervised algorithm in which the noise type is known a priori. One separate basis matrix is obtained for each noise type, and one basis matrix is derived for the speech signal using the training data. A standard NMF in used in the training part, and contribution is made to find a better enhancement procedure using the trained basis matrices. First, we consider a Wiener filtering framework in which the noise PSD is obtained using NMF. Second, we derive a linear minimum mean square error (LMMSE) estimator for the speech signal. Assuming an additive model for the noisy observation, mean square error between the clean speech and the estimated speech components is minimized in the frequency domain to find the estimated speech. The performance of the enhancement algorithms is measured with different instrumental measures including *PESQ*, *SDR*, segmental speech *SNR*, and segmental noise reduction.

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

2. NOTATION AND BASIC CONCEPTS

To refer to the (k,τ)th entry of a matrix X, we use either of notations $X_{k,\tau}$ or $[X]_{k,\tau}$; $\mathbf{x}_\tau$ denotes the τth column of a 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 of the noisy signal. The observation for NMF is obtained by taking the (element-wise) pth power of the magnitude of the DFT coefficients ($X = |Y|^p$). Given the observation matrix X, there are different algorithms to perform the factorization [6, 7]. Here, we use generalized Kullback-Leibler divergence as the cost function:

$$D_{KL}(X\|TV) = \sum_{k,\tau}(X_{k,\tau}\log\frac{X_{k,\tau}}{[TV]_{k,\tau}} + [TV]_{k,\tau} - X_{k,\tau}). \quad (1)$$

Factors T, V are found by iterating the following multiplicative rules [8] to minimize (1):

$$\begin{aligned} T_{k,i} &\leftarrow T_{k,i}\frac{\sum_\tau V_{i,\tau}\left(X_{k,\tau}/[TV]_{k,\tau}\right)}{\sum_p V_{i,p}}, \\ V_{i,\tau} &\leftarrow V_{i,\tau}\frac{\sum_k T_{k,i}\left(X_{k,\tau}/[TV]_{k,\tau}\right)}{\sum_q T_{q,i}}. \end{aligned} \quad (2)$$

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

3. NOISE PSD ESTIMATION USING NMF

In this section, we show how NMF can be used to estimate the noise PSD. The obtained noise PSD will be used in a Wiener filtering framework to enhance the noisy speech. NMF based algorithms consist of training and enhancement phases. 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 observations from the clean speech and noise signals ($|S_{train}|^p$ and $|N_{train}|^p$) to obtain the speech basis matrix, T_S, and noise basis matrix, T_N:

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

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

where S_{train} and N_{train} are the DFT coefficients of the clean speech and noise signals, respectively. Now, the basis matrix for the observed noisy speech, T, is obtained by concatenating T_S and T_N as: $T = (T_S \; T_N)$. In the enhancement phase, an overlap-add framework is utilized to process the noisy speech. Given the vector of the observation at time frame τ (pth power of the magnitude of the DFT coefficients of the τth frame of the noisy speech), $|\mathbf{y}_\tau|^p$, NMF is applied to find a linear approximation of $|\mathbf{y}_\tau|^p$ as: $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau$. In other words, keeping the basis matrix T fixed, NMF is performed to obtain the NMF coefficients vector $\mathbf{u}_\tau$:

$$\mathbf{u}_\tau = \arg\min_{\mathbf{z}} \; D_{KL}(|\mathbf{y}_\tau|^p \,\|T\mathbf{z}). \tag{5}$$

Partitioning $\mathbf{u}_\tau$ as: $\mathbf{u}_\tau = \left(\mathbf{v}_\tau^\top \; \mathbf{w}_\tau^\top\right)^\top$ ($\top$ denotes the transpose), the clean speech component is approximated using $|\mathbf{s}_\tau|^p \approx T_S\mathbf{v}_\tau$, and the noise component is approximated as $|\mathbf{n}_\tau|^p \approx T_N\mathbf{w}_\tau$. An instantaneous estimate of the noise PSD is now obtained as:

$$\widehat{|N_{k,\tau}|^2} = \left(\frac{[T_N\mathbf{w}_\tau]_k}{[T_S\mathbf{v}_\tau + T_N\mathbf{w}_\tau]_k} \times |Y_{k,\tau}|^p\right)^{2/p},$$

Assuming some extent of stationarity of the noise, we can smooth this instantaneous estimate across time to get a better noise PSD estimate:

$$\left[\widehat{\sigma_N^2}\right]_{k,\tau} = \alpha\left[\widehat{\sigma_N^2}\right]_{k,\tau-1} + (1-\alpha)\widehat{|N_{k,\tau}|^2}\;, \tag{6}$$

where $\left[\widehat{\sigma_N^2}\right]_{k,\tau}$ denotes the estimated noise PSD for frequency bin k and time-frame τ.

4. LINEAR MMSE FILTER BASED ON NMF

In this section, we derive a new filter for single channel speech enhancement by minimizing the mean square error between the clean speech and the estimated speech components. We use $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau = T_S\mathbf{v}_\tau + T_N\mathbf{w}_\tau \approx |\mathbf{s}_\tau|^\mathbf{P} + |\mathbf{n}_\tau|^\mathbf{P}$, in which $T_S\mathbf{v}_\tau$ and $T_N\mathbf{w}_\tau$ are some random variables whose specific realizations are to be estimated; Given the observation $|\mathbf{y}_\tau|^p$, we can find the magnitude of the DFT coefficients of the enhanced speech as $\widehat{|S_{k,\tau}|} = \left(\widehat{|S_{k,\tau}|^p}\right)^{1/p}$ where $\widehat{|S_{k,\tau}|^p} = H_{k,\tau}\,|Y_{k,\tau}|^p$ is the linear MMSE estimate of the speech component. Assuming that pth powers of the magnitude of the DFT coefficients at different frequencies are independent, we can minimize the mean square error

$$E\left(\left(|S_{k,\tau}|^p - H_{k,\tau}\,|Y_{k,\tau}|^p\right)^2\right) \approx E\left(\left([T_S\mathbf{v}_\tau]_k - H_{k,\tau}\,[T\mathbf{u}_\tau]_k\right)^2\right) \tag{7}$$

independently for each frequency bin k. H can be obtained by taking the derivative of (7) and making it equal to zero [9, Sec 11.3.1]:

$$0 = \frac{\partial E\left(\left([T_S\mathbf{v}_\tau]_k - H_{k,\tau}[T\mathbf{u}_\tau]_k\right)^2\right)}{\partial H_{k,\tau}} = E\left(-2\,[T_S\mathbf{v}_\tau]_k\,[T\mathbf{u}_\tau]_k + 2H_{k,\tau}\,[T\mathbf{u}_\tau]_k^2\right)$$

and hence:

$$H_{k,\tau} = \frac{E\left([T_S\mathbf{v}_\tau]_k[T\mathbf{u}_\tau]_k\right)}{E\left([T\mathbf{u}_\tau]_k^2\right)}.$$

Assuming independency between the speech and noise components we get:

$$H_{k,\tau} = \frac{E\left([T_S\mathbf{v}_\tau]_k^2\right) + E\left([T_S\mathbf{v}_\tau]_k\right)E\left([T_N\mathbf{w}_\tau]_k\right)}{E\left([T_S\mathbf{v}_\tau]_k^2\right) + E\left([T_N\mathbf{w}_\tau]_k^2\right) + 2E\left([T_S\mathbf{v}_\tau]_k\right)E\left([T_N\mathbf{w}_\tau]_k\right)} \tag{8}$$

Eq. (8) can be converted into a simpler form by assuming that the real and imaginary parts of the DFT coefficients of the speech (S) and noise (N) signals are zero mean normally distributed random variables; recalling that $|\mathbf{s}_\tau|^p \approx T_S\mathbf{v}_\tau$ (and $|\mathbf{n}_\tau|^p \approx T_N\mathbf{w}_\tau$) , the relation between $E\left([T_S\mathbf{v}_\tau]_k^2\right)$ and $E\left([T_S\mathbf{v}_\tau]_k\right)$ (also $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ and $E\left([T_N\mathbf{w}_\tau]_k\right)$) can be found simply for $p = 1, 2$, i.e.:

$$E\left([T_S\mathbf{v}_\tau]_k\right) \approx c\sqrt{E\left([T_S\mathbf{v}_\tau]_k^2\right)}, \tag{9}$$

where $c = \sqrt{\pi}/2$ for $p = 1$, and $c = \sqrt{2}/2$ for $p = 2$.

We can now continue to simplify equation (8). Dividing the denominator and numerator of (8) by $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ and defining $\xi_{k,\tau} = \frac{E\left([T_S\mathbf{v}_\tau]_k^2\right)}{E\left([T_N\mathbf{w}_\tau]_k^2\right)}$, and using (9) we get:

$$H_{k,\tau} \approx \frac{\xi_{k,\tau} + c^2\sqrt{\xi_{k,\tau}}}{\xi_{k,\tau} + 1 + 2c^2\sqrt{\xi_{k,\tau}}}\;, \tag{10}$$

in which we used:

$$\frac{E\left([T_S\mathbf{v}_\tau]_k\right)E\left([T_N\mathbf{w}_\tau]_k\right)}{E\left([T_N\mathbf{w}_\tau]_k^2\right)} \approx c^2\frac{\sqrt{E\left([T_S\mathbf{v}_\tau]_k^2\right)}\sqrt{E\left([T_N\mathbf{w}_\tau]_k^2\right)}}{E\left([T_N\mathbf{w}_\tau]_k^2\right)} = c^2\sqrt{\xi_{k,\tau}}.$$

$\xi_{k,\tau}$ represents the smoothed speech to noise ratio (*smoothed SpNR*). $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ can be estimated by a low pass filter as:

$$E\left([T_N\mathbf{w}_\tau]_k^2\right) \approx \beta E\left([T_N\mathbf{w}_{\tau-1}]_k^2\right) + (1-\beta)\,[T_N\mathbf{w}_\tau]_k^2. \tag{11}$$

$\xi_{k,\tau}$ can be found by following approximation: Define an *approximate SpNR* as $\eta_{k,\tau} = \frac{[T_S\mathbf{v}_\tau]_k^2}{E\left([T_N\mathbf{w}_\tau]_k^2\right)}$, and hence $\xi_{k,\tau} = E(\eta_{k,\tau})$; we propose a decision-directed estimator, similar to [10], for $\xi_{k,\tau}$ as:

$$\xi_{k,\tau} = \max\left(\xi_{min}, \gamma\frac{\widehat{|S_{k,\tau-1}|^p}}{E\left([T_N\mathbf{w}_{\tau-1}]_k^2\right)} + (1-\gamma)\eta_{k,\tau}\right). \tag{12}$$

In our simulations, fairly similar results were obtained using the following approximation of (10):

$$H_{k,\tau} \approx \frac{\xi_{k,\tau}}{\xi_{k,\tau}+1}\;. \tag{13}$$

It is interesting to highlight the differences between (13) and the Wiener filter: Assuming the magnitude of the DFT coefficients of the noisy speech as the observation ($p = 1$) for NMF, and perfect nonnegative factorization for the clean speech and noise signals as $|\mathbf{s}_\tau| = T_S\mathbf{v}_\tau$, $|\mathbf{n}_\tau| = T_N\mathbf{w}_\tau$, (13) will be identical to the Wiener filter; however, by using the magnitude-squared DFT coefficients ($p = 2$) this is not true any more. Moreover, there is another implementation difference between the two filters: the Wiener filter is often implemented by estimating *a priori SNR* based on *a posteriori SNR* which is obtained from the noisy observation [11]; though, for implementing (8), as it is mentioned above, an *approximate SpNR* can be estimated using the separated speech and noise components from NMF; next, the *smoothed SpNR* is estimated using (12) and is used to implement (10) or (13). Since the *approximate SpNR* is based on an initial estimate of the speech component and not the noisy speech, the smoothing factor in (12) should be low enough to capture the speech variations quickly. Good results were obtained for $\gamma = 0.5 - 0.75$ while the results were not sensitive to the exact value of γ. Finally, note that the defined *SpNR* is not the same as the *SNR* which is usually defined as the ratios of the powers of the speech and noise signals.

The algorithm is summarized as:

1. Obtain the NMF coefficients vector $\mathbf{u}_\tau$ by applying NMF to the given observation at time frame τ as $|\mathbf{y}_\tau|^p \approx T\mathbf{u}_\tau$.
2. Find $\mathbf{w}_\tau$ from $\mathbf{u}_\tau = \left(\mathbf{v}_\tau^\top \ \mathbf{w}_\tau^\top\right)^\top$, then obtain an estimate of $E\left([T_N\mathbf{w}_\tau]_k^2\right)$ by smoothing $[T_N\mathbf{w}_\tau]_k^2$ over time (Eq. (11)).
3. Obtain the *approximate SpNR*, $\eta_{k,\tau}$, and *smoothed SpNR*, $\xi_{k,\tau}$ for all frequency bins (Eq. (12)).
4. Obtain the filter gain as (10) or (13).
5. The magnitude of the DFT coefficients of the enhanced speech are obtained as $\widehat{|S_{k,\tau}|} = (H_{k,\tau}\,|Y_{k,\tau}|^p)^{1/p}$.
6. Reconstruct the time domain signal using the noisy phase.

5. EVALUATION

Both magnitude ($p = 1$) and squared magnitude ($p = 2$) of the DFT coefficients of the observed noisy speech signal are used as observation in NMF model for the enhancement task. In the following, the derived algorithm in section 4 (using Eq. (10)) is referred as '*LMMSE-Mag*' and '*LMMSE-Pow*' for $p = 1$ and $p = 2$, respectively. The estimated noise PSDs from Section 3 were used in combination with a Wiener filter to perform the enhancement and are referred as '*Wiener-Mag*' and '*Wiener-Pow*' for $p = 1$ and $p = 2$, respectively; in addition, noise PSD was estimated using a MMSE-based approach [12] which is one of the best algorithms for this purpose [13], and the same Wiener filter was used for the enhancement; in the following, this approach is called '*Wiener-UnS*' to reflect the fact that this approach is an unsupervised filtering and does not have any training. The Wiener filter was implemented using the decision-directed approach [10] with the same parameters $10\log_{10}(\xi_{min}) = -25$dB and $\alpha = 0.98$ for all the approaches. The same lower bound ξ_{min} also was used in (12).

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, and one *speaker independent* model is trained for the speech signal; this model was trained on a mixed group of 24 male and female speakers, and 8 sentences from each speaker were used. For all the approaches 10 sentences form each of the 8 speakers (4 male and 4 female, and none of them were used for the training), and a part of the noise signal which was not used for the training, were used for the performance evaluation. The results are averaged over the entire test set. To apply NMF we use a noise specific basis matrix; if noise type is not known a priori, some adapting procedures have to be used which we have not included in our simulations. For the speech and noise signals, 60 and 100 basis vectors are trained, respectively. The following parameters are obtained by performing a cross-validation test and are used in the simulations: $\alpha = 0.95$ in (6), $\beta = 0.95$ in (11), and $\gamma = 0.6$ in (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 *PESQ* [14], and the Source to Distortion Ratio *(SDR)* which is defined as:

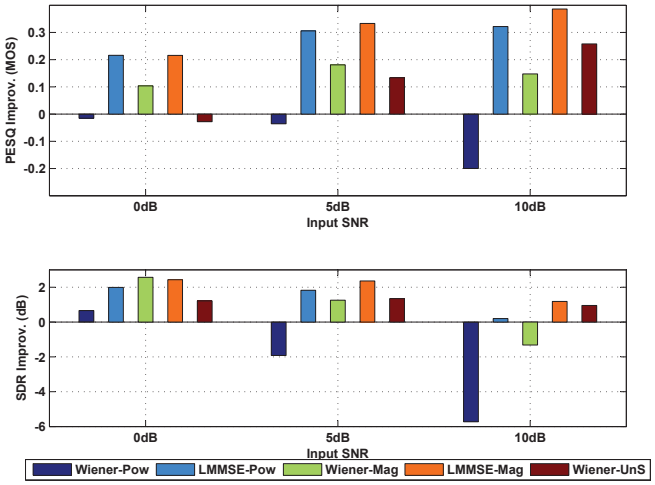
$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{artifact}\|^2},$$

where s_{target}, e_{interf} and $e_{artifact}$ are target time-domain speech signal, interference, and artifact error terms defined in [15], and $\|\cdot\|^2$ denotes the energy. In order to analyze the results more specifically, Segmental speech *SNR* (SNR_{seg-sp}), and segmental noise reduction ($SegNR$) are also measured as [16]:

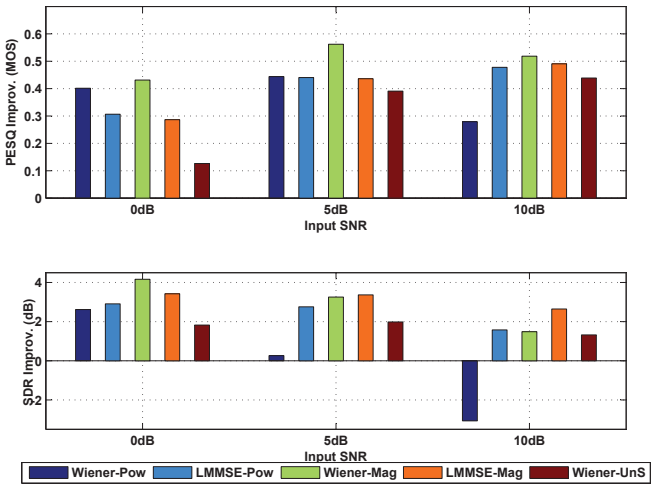
$$SNR_{seg-sp} = \frac{1}{T}\sum_{\tau=1}^{T} 10\log_{10}\left(\frac{\sum_{i=1}^{I} s_{i+\tau I}^2}{\sum_{i=1}^{I}(s_{i+\tau I}-\tilde{s}_{i+\tau I})^2}\right),$$

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

where I denotes the length of the frame, and T the number of frames; These measures are obtained in a shadow filtering framework: the 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.



(a) Babble noise



(b) Factory noise

Figure 1: *PESQ* and *SDR* improvements for babble and factory noises.

5.1. Results and Discussion

Figure 1 shows the improvement in *PESQ* and *SDR* for different algorithms. The results show that a NMF-based filter which is derived using the magnitude of the DFT coefficients ($p = 1$) of the noisy speech gives a better result compared to the same type of the NMF-based filter which is derived using the magnitude-squared

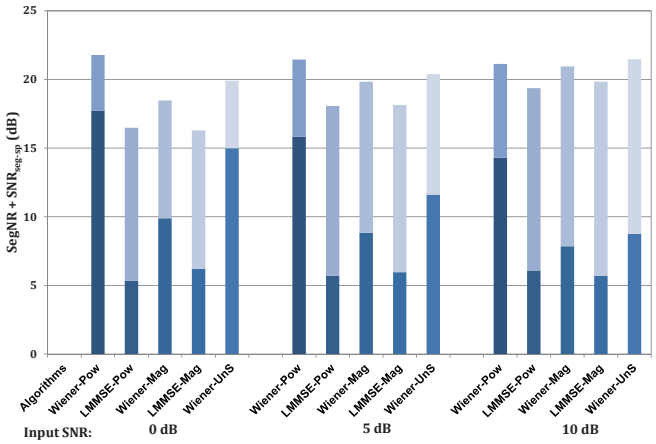


Figure 2: Stacked presentation of Segmental Noise Reduction ($SegNR$, bottom), and Segmental Speech *SNR* (SNR_{seg-sp}, top) for factory noise.

DFT coefficients ($p = 2$). This is true for both the Wiener and the proposed LMMSE filters, and agrees with the previous applications of NMF in source separation (e.g. [1, 2]). For the Wiener based algorithms, the difference between these two cases is much higher than that between the LMMSE algorithms. The NMF-based algorithms with $p = 1$ mostly result to a better performance than the *Wiener-UnS* algorithm, especially at low input *SNR*s. In the most cases the *SDR* improvement for *LMMSE-Mag* is the highest among all the algorithms for both noise types. For the babble noise, the *PESQ* improvement is higher for the proposed LMMSE algorithms compared to the Wiener based algorithms. For the factory noise, the Wiener based algorithms often provide better *PESQ* for the enhanced speech signal than the LMMSE algorithms, especially for low input *SNR*s.

For input *SNR*s for which *PESQ* and *SDR* improvements are not pointing in the same direction (for instance factory noise at 5 dB input *SNR*) it becomes more difficult to compare different algorithms; hence, we performed an informal listening test and found that if the difference in the *PESQ* improvements is not high, and at the same time the difference in the *SDR* improvements is high, the algorithm with the higher *SDR* is preferred; for example, the *LMMSE-Pow* was preferred over the *Wiener-Pow* algorithm for factory noise at 5 dB input *SNR*. This is because *LMMSE-Pow* provides much higher *SDR* even though both methods provide similar *PESQ* scores for the enhanced speech. This can be expected since none of these measures completely model the speech quality. Even fairly similar scores were obtained for the *LMMSE-Mag* and *Wiener-Mag* for the factory noise at 5 dB input *SNR*. These results might be explained by looking at Figure 2.

Figure 2 shows the stacked results for Segmental Noise Reduction, $SegNR$, and Segmental Speech *SNR*, SNR_{seg-sp}, for factory noise which are shown in the bottom and top of the figure respectively. For both measures a high value is desired. SNR_{seg-sp} is inversely proportional to the speech distortion. Wiener based approaches provide a higher $SegNR$ and lower SNR_{seg-sp} compared to the proposed LMMSE algorithms. The *PESQ* improvements for the Wiener based approaches are obtained mainly because of the high $SegNR$ while for the proposed LMMSE filters the *PESQ* improvements are obtained mainly because of the high SNR_{seg-sp} (and hence less speech distortion).

6. CONCLUSIONS

Two types of NMF-based algorithms were obtained in this paper: first, a Wiener filter was considered in which noise PSD was estimated using NMF. Second, a LMMSE filter was derived by minimizing the mean square error between the clean speech and the estimated speech components in the frequency domain. The proposed LMMSE filters were shown to be promising and gave better *SDR* improvements compared to the Wiener-based algorithms in most of the test cases; LMMSE filters gave a higher *PESQ* improvements for the babble noise for all the simulated input *SNR*s although for the factory noise it was not the case. Most of the NMF-based approaches gave better *SDR* and *PESQ* improvements compared to the Wiener filtering method in which a recently developed unsupervised approach was used to estimate the noise PSD.

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

[4] 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.

[5] 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.

[6] 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.

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

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

[9] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, 2006.

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

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

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

[13] 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.

[14] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrow-band telephone networks and speech codecs," Tech. Rep., 2000.

[15] 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.

[16] 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.