Speech Communication 87 (2017) 18–30

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# Regularized non-negative matrix factorization with Gaussian mixtures and masking model for speech enhancement[☆]

Hanwook Chung[a,*], Eric Plourde[b], Benoit Champagne[a]

[a] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada
[b] Department of Electrical and Computer Engineering, Sherbrooke University, Sherbrooke, Quebec, Canada

ARTICLE INFO

*Article history:*
Received 4 September 2015
Available online 11 November 2016

*Keywords:*
Single-channel speech enhancement
Supervised algorithm
Non-negative matrix factorization
Regularization
Gaussian mixture model
Masking effects

ABSTRACT

We introduce single-channel supervised speech enhancement algorithms based on regularized non-negative matrix factorization (NMF). In the proposed framework, the log-likelihood functions (LLF) of the magnitude spectra for both the clean speech and noise, based on Gaussian mixture models (GMM), are included as regularization terms in the NMF cost function. By using this proposed regularization as *a priori* information in the enhancement stage, we can exploit the statistical properties of both the clean speech and noise signals. For further improvement of the enhanced speech quality, we also incorporate a masking model of the human auditory system in our approach. Specifically, we construct a weighted Wiener filter (WWF) where the power spectral densities (PSD) of the speech and noise are estimated from the above mentioned NMF algorithm with the proposed regularization. The weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the enhanced speech. Experimental results of perceptual evaluation of speech quality (PESQ), source-to-distortion ratio (SDR) and segmental signal-to-noise ratio (SNR) show that the proposed speech enhancement algorithms (i.e., regularized NMF with and without masking model) provide better performance in speech enhancement than the benchmark algorithms.

## 1. Introduction

Speech enhancement algorithms aim to remove additive background noise from a noisy speech signal in order to improve its quality or intelligibility. They have been an attractive research area for decades and find diverse applications, including mobile telephony, hearing aid and speech recognition, to name a few. Numerous algorithms for single channel speech enhancement have been proposed in the past, such as: Wiener filtering (Lim and Oppenheim, 1979; Scalart and Filho, 1996), spectral subtraction (Boll, 1979; Virag, 1999), minimum mean-square error (MMSE) estimation of the short-time spectral amplitude (STSA) (Ephraim and Malah, 1984; Loizou, 2005; Plourde and Champagne, 2008; You et al., 2005) and subspace decomposition (Ephraim and Van Trees, 1995; Hermus et al., 2007; Jensen et al., 1995). However, these algorithms use a minimal amount of *a priori* information about the speech and noise. Consequently, they tend to provide limited performance gains, especially when the speech is contaminated by adverse noise, such as under low signal-to-noise ratio (SNR) or non-stationary noise conditions.

Further improvements of the MMSE-based estimators have been proposed by modeling the speech spectrum as a Rayleigh mixture model (RMM) (Erkelens et al., 2007) or a Gaussian mixture model (GMM) (Ding et al., 2005; Hao et al., 2010). These estimators, which use model parameters derived from a training set for the clean speech, provide a more detailed and accurate description of the speech distribution and are better suited to handle non-stationary speech features. In contrast to the speech model, the parameters of the noise distribution are often estimated directly from the noisy speech spectrum. These can be obtained by using an estimation algorithm where the noise power spectral density (PSD) is calculated recursively over successive time frames to capture non-stationary features (Cohen, 2003; Gerkmann and Hendriks, 2012; Rangachari and Loizou, 2006). However, the noise spectrum is modeled by a single distribution which is one of the main limitations of the above MMSE-based estimators.

Recently, the non-negative matrix factorization (NMF) approach has been applied to various problems such as image representation (Zafeiriou et al., 2006), music transcription (Bertin et al., 2010), source separation (Virtanen, 2007a) and speech

[☆] Funding for this work was provided by Microsemi Corporation (Ottawa, Canada) and a grant from NSERC (Govt. of Canada)

[*] Corresponding author.
*E-mail addresses:* hanwook.chung@mail.mcgill.ca (H. Chung), eric.plourde@usherbrooke.ca (E. Plourde), benoit.champagne@mcgill.ca (B. Champagne).

http://dx.doi.org/10.1016/j.specom.2016.11.003
0167-6393/© 2016 Elsevier B.V. All rights reserved.

enhancement (Mohammadiha et al., 2013). In general, NMF is a dimensionality reduction tool, which decomposes a given data matrix into basis and activation matrices with non-negative elements constraint (Févotte et al., 2009; Lee and Seung, 2001). In speech and audio applications, the magnitude or power spectrum of the desired signal is interpreted as a linear combination of the basis vectors. In supervised learning-based NMF algorithms, the basis vectors are obtained for each source independently by employing training data, and subsequently used during the separation or enhancement stage (Grais and Erdogan, 2013; Mohammadiha et al., 2013). However, one of the main problems of such supervised algorithms is the existence of a mismatch between the characteristics of the training and test data, which in turn leads to a decreased quality of the estimated source signals. One possible remedy to this problem is to add explicit regularization terms to the NMF cost function that incorporate some prior knowledge. In order to account for the temporal dependency of the successive time frames, Févotte et al. (2009) model the activations by means of Markov chain, Grais and Erdoğan (2012) and Mysore and Smaragdis (2011) use a hidden Markov model (HMM), while Grais and Erdogan (2013) use GMMs that help the activations to follow certain patterns. In Chung et al. (2014), both the speech and noise spectra are modeled by a GMM, and their log-likelihood functions (LLF) are used as regularization terms.

Besides the speech enhancement or source separation algorithms which mainly focus on the perspective of signal estimation and reconstruction, several algorithms incorporating modeling aspects of the human auditory system have been proposed in order to improve the perceptual quality of the estimated source signals. Specifically, these refined algorithms exploit a psychoacoustical property called auditory masking which refers to a process whereby one sound is rendered inaudible due to the presence of another sound (Fastl and Zwicker, 2007). In the case of frequency domain (or simultaneous) masking, the threshold which models this effect has been used for selecting parameters in spectral subtraction (Virag, 1999), subspace decomposition (Jabloun and Champagne, 2003), Wiener filtering (Hu and Loizou, 2004) and MMSE-based estimator (Hansen et al., 2006; Natarajan et al., 2005). In the NMF-based algorithms, weighted NMF update rules have been proposed by applying a weighting matrix based on the masking threshold to the NMF cost function (Kırbız and Günsel, 2013; Virtanen, 2007b). For speech enhancement, the masking threshold which determines the amount of the noise reduction is usually calculated from the estimated PSD of the clean speech. This suggests that a more accurate estimation scheme may lead to further improvement of the enhanced speech quality when applying a masking threshold.

In this paper, we introduce single-channel supervised speech enhancement algorithms based on regularized NMF which are extensions of our previous work (Chung et al., 2014). The proposed framework seeks to exploit the statistical properties of *both* the clean speech and noise, an approach which is widely used in traditional speech enhancement algorithms. This is achieved in two ways: i) by representing the corresponding magnitude spectra, which capture the general (*high-level*) characteristics of the signals, with the help of GMMs motivated by Ding et al. (2005) and Hao et al. (2010), and ii) by adding regularization terms that incorporate this *a priori* information to the NMF cost function in the enhancement stage. The proposed method, therefore, can be interpreted as a combination of the NMF and statistical model-based approaches. During the training stage, by using an isolated training set for each type of clean speech and noise, we estimate the basis matrices in the NMF model via multiplicative update rules (Lee and Seung, 2001) and the parameters of the GMMs via the expectation-maximization (EM) algorithm (Bishop, 2006; Dempster et al., 1977). For the GMM, we propose to use normalized spectral values in order to handle the magnitude difference between the training and test data, similar to the work of Grais and Erdogan (2013). In the enhancement stage, the LLFs of the clean speech and noise magnitude spectra are added as regularization terms to the NMF cost function and the activation matrix of the noisy speech is estimated. Consequently, the PSDs of the clean speech and noise are obtained and the enhanced speech is reconstructed using Wiener filtering.

For further improvement of the enhanced speech quality, we incorporate the masking effects of the human auditory system in our approach. Specifically, we construct a weighted Wiener filter (WWF) where the PSDs of the speech and noise are estimated from the above mentioned NMF algorithm with the proposed regularization. The weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the speech based on Painter and Spanias (2000). Experimental results of perceptual evaluation of speech quality (PESQ) (Recommendation, 2001), source-to-distortion ratio (Vincent et al., 2006) and segmental signal-to-noise ratio (SNR) show that the proposed speech enhancement algorithms provide better performance in speech enhancement than the benchmark algorithms.

The rest of the paper is organized as follows. In Section 2, we briefly review the basic principles of NMF-based single channel speech enhancement. The proposed NMF training stage with GMM parameter estimation is described in Section 3. In Section 4, the proposed modifications to the enhancement stage, including NMF algorithm with regularization, masking threshold estimation and perceptually motivated NMF algorithm for speech enhancement are explained. Experimental results are presented in Section 5 and finally, a conclusion is given in Section 6.

## 2. NMF-based speech enhancement

For a given matrix $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K\times L}$, NMF finds a local optimal decomposition $\mathbf{V} = \mathbf{WH}$, where $\mathbf{W} = [w_{km}] \in \mathbb{R}_+^{K\times M}$ is a basis matrix, $\mathbf{H} = [h_{ml}] \in \mathbb{R}_+^{M\times L}$ is an activation matrix, $\mathbb{R}_+$ denotes the set of non-negative real numbers and $M$ is the number of basis vectors, typically chosen such that $KM + ML \ll KL$ (Févotte et al., 2009; Lee and Seung, 2001). The factorization is obtained by minimizing a suitable cost function, denoted as $\mathcal{J}(\mathbf{V}, \mathbf{WH})$. By expressing the gradient of the cost function as the difference of two non-negative terms such that $\nabla\mathcal{J}(\mathbf{V}, \mathbf{WH}) = \nabla^+\mathcal{J}(\mathbf{V}, \mathbf{WH}) - \nabla^-\mathcal{J}(\mathbf{V}, \mathbf{WH})$, solutions can be obtained iteratively using the following heuristic multiplicative update rules (Bertin et al., 2010; Févotte et al., 2009; Grais and Erdogan, 2013):

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{\nabla_{\mathbf{W}}^- \mathcal{J}(\mathbf{V}, \mathbf{WH})}{\nabla_{\mathbf{W}}^+ \mathcal{J}(\mathbf{V}, \mathbf{WH})}, \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\nabla_{\mathbf{H}}^- \mathcal{J}(\mathbf{V}, \mathbf{WH})}{\nabla_{\mathbf{H}}^+ \mathcal{J}(\mathbf{V}, \mathbf{WH})} \tag{1}$$

where the operator $\otimes$ and the quotient line respectively denote element-wise multiplication and division, and the $\leftarrow$ refers to an iterative overwrite. Among various cost functions, the most widely used one is the Kullback–Leibler (KL) divergence (e.g., FitzGerald et al., 2008), defined as

$$\mathcal{J}(\mathbf{V}, \mathbf{WH}) = \mathcal{D}_{KL}(\mathbf{V}, \mathbf{WH}) \triangleq \sum_{k=1}^{K}\sum_{l=1}^{L}\left(v_{kl}\ln\frac{v_{kl}}{[\mathbf{WH}]_{kl}} - v_{kl} + [\mathbf{WH}]_{kl}\right) \tag{2}$$

where $[\,\cdot\,]_{kl}$ denotes the $(k, l)$th entry of its matrix argument. The update rules of the NMF with KL-divergence based on (1) are given as

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{(\mathbf{V}/(\mathbf{WH}))\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T}, \qquad \mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T(\mathbf{V}/(\mathbf{WH}))}{\mathbf{W}^T\mathbf{1}} \tag{3}$$

where $\mathbf{1}$ is a $K \times L$ matrix with all entries equal to one, the operator / denotes element-wise division and the superscript $T$ denotes matrix transpose. The scale indeterminacies of $\mathbf{W}$ and $\mathbf{H}$ can be prevented by including a normalization step which leaves the cost function unchanged (Févotte et al., 2009). Specifically, at the end of each iteration, we can use the $l_1$-norm to normalize the column vectors of the basis matrix, $\mathbf{W}$, and scale the row vectors of the activation matrix, $\mathbf{H}$, accordingly, e.g., Cichocki et al. (2006); Zafeiriou et al. (2006). As for the initialization of $\mathbf{W}$ and $\mathbf{H}$, positive random numbers are commonly used (Févotte et al., 2009). Numerical instability due to division by zero or taking the logarithm of zero, which may appear in the KL-divergence in (2) or in the update rules given by (3), can be avoided in a practical implementation by adding a small positive number, e.g., $10^{-20}$, to the various denominators in (2) and (3) and the numerator of the log function in (3), e.g., (Cichocki et al., 2006; Lefevre et al., 2011).

Note that the update rules given in (1) do not guarantee the convergence to a stationary point in general (Févotte et al., 2009). Nevertheless, they are widely used due to the simplicity of their derivation and implementation, especially in diverse regularized algorithms, e.g., Virtanen (2007a), Grais and Erdogan (2013). By adding an additional regularization term to the KL-divergence, we can construct a regularized cost function as,

$$\mathcal{J}(\mathbf{V},\mathbf{WH}) = \mathcal{D}_{KL}(\mathbf{V},\mathbf{WH}) + \alpha\mathcal{R}(\mathbf{W},\mathbf{H}) \tag{4}$$

where $\alpha > 0$ is a regularization coefficient and $\mathcal{R}(\mathbf{W},\mathbf{H})$ denotes a regularization term. An iterative solution algorithm is easily obtained using the update rules given in (1). Various approaches for choosing the regularization term have been introduced by considering sparsity (Virtanen, 2007a), temporal continuity (Bertin et al., 2010; Virtanen, 2007a), harmonicity of music signals (Bertin et al., 2010) and statistical priors (Chung et al., 2014; Grais and Erdogan, 2013).

In single-channel speech enhancement, the observed noisy speech signal can be expressed in the time-frequency domain via the short-time Fourier transform (STFT) as (O'Shaughnessy, 1987),

$$Y(k,l) = S(k,l) + N(k,l) \tag{5}$$

where $Y(k, l)$, $S(k, l)$ and $N(k, l)$ respectively denote the STFT of the noisy speech, clean speech and noise for the $k$th frequency bin of the $l$th time frame. We assume that the magnitude spectrum of the noisy speech can be approximated by $|Y(k,l)| \approx |S(k,l)| + |N(k,l)|$, as it is a practical assumption widely used in NMF-based audio and speech signal processing (Grais and Erdogan, 2013; Mohammadiha et al., 2011; Virtanen, 2007a). Throughout this paper, we will use the following notations to represent the magnitude spectrum matrices of the different signals under consideration: $\mathbf{V} = [v_{kl}] \in \mathbb{R}_+^{K\times L}$ where $v_{kl}$ is the magnitude spectral value for the $k$th frequency bin of the $l$th time frame, $K$ is the number of frequency bins and $L$ is the number of time frames. Furthermore, we shall use the subscripts or superscripts $Y$, $S$ and $N$, respectively, to indicate the noisy speech, clean speech and noise (as in, e.g., $v_{kl}^Y = |Y(k,l)|$). We also adopt a similar convention for the basis and activation matrices.

In general, NMF-based supervised speech enhancement algorithms consist of two stages (Grais and Erdogan, 2013; Mohammadiha et al., 2013). During the training stage, by applying (3) to the training data $\mathbf{V}_S \in \mathbb{R}_+^{K\times L_S}$ and $\mathbf{V}_N \in \mathbb{R}_+^{K\times L_N}$ separately, the basis matrices for both the clean speech and noise, $\mathbf{W}_S = [w_{km}^S] \in \mathbb{R}_+^{K\times M_S}$ and $\mathbf{W}_N = [w_{km}^N] \in \mathbb{R}_+^{K\times M_N}$, are obtained. The activation matrices for the clean speech and noise, which are computed along with the basis matrices, are discarded after the training stage. In the enhancement stage, by fixing these basis matrices as $\mathbf{W}_Y = [\mathbf{W}_S \;\; \mathbf{W}_N] \in \mathbb{R}_+^{K\times(M_S+M_N)}$, the activation matrix of the noisy speech is estimated, i.e., $\hat{\mathbf{H}}_Y = [\hat{\mathbf{H}}_S^T \;\; \hat{\mathbf{H}}_N^T]^T \in \mathbb{R}_+^{(M_S+M_N)\times L_Y}$, by applying the NMF activation update in (3) to the noisy speech magnitude spectrum $\mathbf{V}_Y \in \mathbb{R}_+^{K\times L_Y}$. Note that the regularized NMF algorithm can be applied instead to exploit some prior knowledge of the signals, where the update rules can be derived by using the heuristic multiplicative update rules given in (1) based on the cost function given in (4). Once the activation matrix of the noisy speech is obtained, the clean speech spectrum can be estimated using a Wiener filter (WF) as (Févotte et al., 2009; Kırbız and Günsel, 2013; Mohammadiha et al., 2011),

$$\hat{\mathbf{S}} = \frac{\hat{\mathbf{P}}_S}{\hat{\mathbf{P}}_S + \hat{\mathbf{P}}_N} \otimes \mathbf{Y} \tag{6}$$

where $\hat{\mathbf{P}}_S = [\hat{P}_S(k,l)]$ and $\hat{\mathbf{P}}_N = [\hat{P}_N(k,l)] \in \mathbb{R}_+^{K\times L_Y}$ respectively denote the estimated power spectral density (PSD) matrices of the clean speech and noise and $\mathbf{Y} = [Y(k,l)] \in \mathbb{C}^{K\times L_Y}$ denotes the matrix of noisy speech STFT coefficients. Hence, the estimated clean speech in (6) makes use of the phase from the initial noisy speech in $\mathbf{Y}$.[1]The PSDs can be obtained via temporal smoothing of the NMF-based periodograms as given by Kwon et al. (2015),

$$\hat{P}_S(k,l) = \tau_S\hat{P}_S(k,l-1) + (1-\tau_S)([\mathbf{W}_S\hat{\mathbf{H}}_S]_{kl})^2 \tag{7}$$

$$\hat{P}_N(k,l) = \tau_N\hat{P}_N(k,l-1) + (1-\tau_N)([\mathbf{W}_N\hat{\mathbf{H}}_N]_{kl})^2 \tag{8}$$

where $\tau_S$ and $\tau_N$ are the temporal smoothing factors for the speech and noise, respectively. Finally, the enhanced speech signal in the time-domain is reconstructed by applying an inverse STFT on (6) followed by the overlap-add method (O'Shaughnessy, 1987).

## 3. Proposed training stage

In the proposed framework, *a priori* knowledge about the magnitude spectra of the clean speech and noise is captured by distinct GMMs. As a brief overview of the training stage, we first estimate the basis and activation matrices for the clean speech and noise independently using isolated training data. To this end, we consider the KL-divergence given in (2) and apply the resulting update rules in (3), leading to factorizations $\mathbf{V}_S = \mathbf{W}_S\mathbf{H}_S$ and $\mathbf{V}_N = \mathbf{W}_N\mathbf{H}_N$. Subsequently, the GMM parameters for the speech and noise are estimated from the corresponding NMF parameters. The details of this computation, which is identical for the speech and noise, are further developed below where for convenience in notation, the subscripts $S$ and $N$ are dropped.

In Ding et al. (2005) and Hao et al. (2010), the probability density function (PDF) of the clean speech spectrum is modeled by a GMM. Motivated by this approach, we model the PDFs of the magnitude spectra for *both* the clean speech and noise by distinct GMMs.[2] Therefore, we can expect that a more detailed and accurate statistical description is provided for the noise as well as the clean speech. In the proposed algorithm, we consider the product $\mathbf{WH}$, which is an approximation of $\mathbf{V}$, as the observation matrix for the parameter estimation of the magnitude spectrum PDF,[3] since we intend to introduce a clear connection with the regularization term shown in (4). Specifically, by expressing the observation as $\mathbf{WH}$, we can directly differentiate the regularization term with respect to $\mathbf{H}$ while deriving the update rule given by (1) during the enhancement stage (a detailed derivation will be presented in Section 4.1). Moreover, in order to handle the magnitude difference between the training and test data, we consider normalized

[1] According to (6), only the magnitude of the noisy speech $Y(k, l)$ is modified during the enhancement stage. This approach is common in most of the literature on speech enhancement (O'Shaughnessy, 1987).

[2] Alternatively, we can model the PDF of the magnitude spectra by a RMM (e.g., Erkelens et al., 2007) or Gamma mixture model (e.g., Virtanen and Cemgil, 2009), which remain an interesting avenue for our future explorative work.

[3] Indeed, we could verify through independent experiments that there was no significant difference in the enhancement performance when considering either $\mathbf{V}$ or $\mathbf{WH}$ as the observation matrix.

observations where the columns of $\mathbf{WH}$ are normalized by their $l_1$-norm,[4] similar to Grais and Erdogan (2013). Specifically, we define the normalized column of the observation matrix as,

$$\bar{\mathbf{V}}_l \triangleq \frac{[\mathbf{WH}]_l}{\sum_m h_{ml}} \tag{9}$$

where $[\cdot]_l$ denotes the $l$th column of its matrix argument. Note that the $l_1$-norm of $[\mathbf{WH}]_l$, i.e., $\sum_k [\mathbf{WH}]_{kl}$, simply turns into $\Sigma_m h_{ml}$ since the basis vectors are normalized with respect to the $l_1$-norm, i.e., $\sum_k w_{km} = 1$ for $m \in \{1, ...M\}$. The GMM is defined in terms of the following parametric model for the PDF of $\bar{\mathbf{V}}_l$

$$p(\bar{\mathbf{V}}_l|\boldsymbol{\theta}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\bar{\mathbf{V}}_l|\mathbf{z}) = \sum_{i=1}^{I} g_i \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{10}$$

where $I$ is the number of Gaussian components, $\mathbf{z} = [z_1, ..., z_I]^T$ is an $I$-dimensional vector of discrete latent variables $z_i \in \{0, 1\}$ with $\sum_i z_i = 1$, and the set $\boldsymbol{\theta} \triangleq \{g_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{I}$ consists of the GMM parameters. The marginal distribution over $\mathbf{z}$ is specified in terms of the mixing coefficients $g_i \triangleq p(z_i = 1)$. The conditional PDF of $\bar{\mathbf{V}}_l$ given a particular value for the latent variable $z_i$ is a $K$-dimensional Gaussian distribution such that $p(\bar{\mathbf{V}}_l|z_i = 1) = \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where $\boldsymbol{\mu}_i = [\mu_{i,k}]$ is the mean vector and $\boldsymbol{\Sigma}_i$ is the covariance matrix. In this work, we ignore possible correlations between different spectral components and therefore consider diagonal covariance matrices for simplicity, i.e., $\boldsymbol{\Sigma}_i = \text{diag}\{\sigma_{i,k}^2\}$. Recall that the entries of the observation matrix $\bar{\mathbf{V}} = [\bar{v}_{kl}]$ are magnitude spectral values which are strictly non-negative, while the GMM can in theory assign non-zero probability to negative values. Nevertheless, modeling matrix $\bar{\mathbf{V}}$ by a GMM is perfectly reasonable if the mean value of its entries exceed the corresponding standard deviation by a significant margin. More specifically, if say $\mu_{i,\,k} \geq 3\sigma_{i,\,k}$ for every Gaussian component $i = 1, ..., I$, then we can safely assume that $P_r[\bar{v}_{kl} < 0] \approx 0$. In effect, we have been able to verify that this condition is generally satisfied in our experimental work.

The parameter set $\boldsymbol{\theta} = \{g_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{I}$ can be estimated using the expectation-maximization (EM) algorithm (Bishop, 2006; Dempster et al., 1977). For a given observation $\bar{\mathbf{V}} = [\bar{\mathbf{V}}_1, \bar{\mathbf{V}}_2, ..., \bar{\mathbf{V}}_L] = [\bar{v}_{kl}]$, where the column vectors $\bar{\mathbf{V}}_l$ are assumed to be drawn independently, the LLF can be written as,

$$\begin{aligned}\mathcal{L}(\bar{\mathbf{V}}|\boldsymbol{\theta}) &\triangleq \ln p(\bar{\mathbf{V}}|\boldsymbol{\theta}) \\ &= \sum_{l=1}^{L} \ln\left\{\sum_{i=1}^{I} g_i \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right\} \\ &\geq \sum_{l=1}^{L}\sum_{i=1}^{I} q(z_i) \ln\left\{\frac{g_i \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{q(z_i)}\right\} \triangleq \mathcal{L}_B(\bar{\mathbf{V}}|\boldsymbol{\theta})\end{aligned} \tag{11}$$

where $q(z_i)$ is an arbitrary probability distribution. The inequality holds for any choice of $q(z_i)$ due to Jensen's inequality (Cemgil, 2009; Hao et al., 2010). Note that $\mathcal{L}_B(\bar{\mathbf{V}}|\boldsymbol{\theta})$ defines a lower bound on $\mathcal{L}(\bar{\mathbf{V}}|\boldsymbol{\theta})$ where the equality holds for $q(z_i) = p(z_i = 1|\bar{\mathbf{V}}_l, \boldsymbol{\theta})$, which is the posterior distribution of latent variable $z_i$ given the observation $\bar{\mathbf{V}}_l$. The EM algorithm is an iterative procedure which consists of two steps. During the expectation step (E-step), the posterior distribution of each latent variable given the observation is calculated, which is shown as

$$\gamma_{il}^{(r)} \triangleq p(z_i = 1|\bar{\mathbf{V}}_l, \boldsymbol{\theta}^{(r)}) = \frac{g_i^{(r)} \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i^{(r)}, \boldsymbol{\Sigma}_i^{(r)})}{\sum_{i=1}^{I} g_i^{(r)} \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i^{(r)}, \boldsymbol{\Sigma}_i^{(r)})} \tag{12}$$

where the superscript $(r)$ denotes the $r$th iteration. In the maximization step (M-step), by *fixing* the posterior distribution to $\gamma_{il}^{(r)}$, the parameter set $\boldsymbol{\theta}$ which maximizes $\mathcal{L}_B(\bar{\mathbf{V}}|\boldsymbol{\theta})$ is determined. In effect, since $\gamma_{il}^{(r)}$ in (12) does not depend on $\boldsymbol{\theta}$, this is equivalent to the maximization criterion of the expectation of the complete data LLF with respect to the posterior distribution,

$$\mathcal{L}_C(\bar{\mathbf{V}}|\boldsymbol{\theta}) \triangleq \sum_{l=1}^{L}\sum_{i=1}^{I} \gamma_{il}^{(r)} \ln\{g_i \mathcal{N}(\bar{\mathbf{V}}_l|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}. \tag{13}$$

The solution of the M-step can be obtained in closed form as,

$$\begin{aligned} g_i^{(r+1)} &= \frac{1}{L}\sum_{l=1}^{L} \gamma_{il}^{(r)}, \\ \mu_{i,k}^{(r+1)} &= \frac{\sum_{l=1}^{L} \gamma_{il}^{(r)} \bar{v}_{kl}}{\sum_{l=1}^{L} \gamma_{il}^{(r)}}, \\ \sigma_{i,k}^{2\ \ (r+1)} &= \frac{\sum_{l=1}^{L} \gamma_{il}^{(r)} (\bar{v}_{kl} - \mu_{i,k}^{(r+1)})^2}{\sum_{l=1}^{L} \gamma_{il}^{(r)}}. \end{aligned} \tag{14}$$

As for the initialization of $\boldsymbol{\theta}$, we apply $k$-means clustering to $\bar{\mathbf{V}}$, which is an iterative algorithm aiming to partition the observations into clusters, such that each observation belongs to the cluster with the nearest mean (Bishop, 2006). The number of clusters is set equal to $I$, the number of Gaussian components in the GMM, while the cluster mean values are initialized randomly.

At this point, we emphasize the main difference between the above proposed training algorithm and the one presented in our previous work (Chung et al., 2014). In the latter, we considered joint training of $\mathbf{W}$, $\mathbf{H}$ and $\boldsymbol{\theta}$, where we used a regularized cost function as in (4) in which the regularization term was the expected LLF given in (13). We observed that the regularization coefficient $\alpha$ not only determines the convergence behavior of the iterative update but that it also affects the enhancement performance. Hence, selecting an appropriate value for this coefficient is difficult. In addition, the iterative update using the joint training converges slowly and hence requires a more extensive computational effort. For these reasons, we chose to consider here instead a *sequential* form of training, which is found to be simpler and more efficient in both terms of computation and enhancement performance.

## 4. Proposed enhancement stage

In this section, we introduce the proposed regularized NMF algorithms. The LLF of the magnitude spectra for both the clean speech and noise based on distinct GMMs are included as regularization terms in the NMF cost function, which will be discussed in Section 4.1. For further improvement of enhancement performance, we incorporate a masking model of the human auditory system in our approach, which will be provided in Section 4.2. Specifically, we construct a WWF where the PSDs of the speech and noise are estimated by using the method in Section 4.1, and the weighting factor in the WWF is selected based on a masking threshold which is obtained from the estimated PSD of the clean speech.

### 4.1. Regularized NMF with Gaussian mixtures

In the proposed enhancement stage, the activation matrix of the noisy speech $\mathbf{H}_Y = [\mathbf{H}_S^T\ \mathbf{H}_N^T]^T$ is estimated using the regularized NMF algorithm based on (1) and (4), by fixing the basis matrices $\mathbf{W}_Y = [\mathbf{W}_S\ \mathbf{W}_N]$ and the GMM parameter sets of the clean speech and noise, $\boldsymbol{\theta}_S = \{g_i^S, \boldsymbol{\mu}_i^S, \boldsymbol{\Sigma}_i^S\}_{i=1}^{I_S}$ and $\boldsymbol{\theta}_N = \{g_i^N, \boldsymbol{\mu}_i^N, \boldsymbol{\Sigma}_i^N\}_{i=1}^{I_N}$, which are obtained during the training stage. Specifically, the LLFs of the clean speech and noise based on (11), i.e., $\mathcal{L}(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S)$ and $\mathcal{L}(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_N)$, are used as regularization terms. The proposed regularized cost function is shown as,

$$\mathcal{J} = \mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y) - \mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) \tag{15}$$

[4] Note that this normalization step differs from the one included in the NMF update introduced in Section 2, where we normalize the basis matrix and scale the activation matrix accordingly to avoid the scale indeterminacy.

where $\mathcal{D}_{KL}(\cdot)$ is the KL-divergence given in (2) and $\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ is the proposed regularization term written as,

$$\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \alpha_S \mathcal{L}(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S) + \alpha_N \mathcal{L}(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_N) \tag{16}$$

where $\mathcal{L}(\cdot|\cdot)$ is given in (11) and $\bar{\mathbf{V}}_S$, $\bar{\mathbf{V}}_N$ are the normalized clean speech and noise spectra defined by (9). The values $\alpha_S > 0$ and $\alpha_N > 0$ are the regularization coefficients for the clean speech and noise, respectively. The optimal choices for $\alpha_S$ and $\alpha_N$ depend on the input SNR as well as the speaker, the type of noise and regularization term. In this paper, however, we do not consider such dependencies (except the type of regularization term), and use constant values for simplicity, as we found indeed that the optimal choices mostly depend on the regularization term. Note that a negative sign is applied to the regularization term in (15), since the latter will represent a reward as opposed to a penalty.

For the derivation of the update rule of $\mathbf{H}_Y$, we first compute the gradient of $\mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y)$ with respect to $\mathbf{H}_Y$. This gradient is shown as

$$\nabla_{\mathbf{H}_Y}\mathcal{D}_{KL} = \nabla^+_{\mathbf{H}_Y}\mathcal{D}_{KL} - \nabla^-_{\mathbf{H}_Y}\mathcal{D}_{KL} \tag{17}$$

where the dependence of $\mathcal{D}_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y)$ on $\mathbf{V}_Y$ and $\mathbf{W}_Y\mathbf{H}_Y$ is omitted for notational convenience, and the values on the right-hand side are

$$\nabla^+_{\mathbf{H}_Y}\mathcal{D}_{KL} = \mathbf{W}_Y^T\mathbf{1} \tag{18}$$

$$\nabla^-_{\mathbf{H}_Y}\mathcal{D}_{KL} = \mathbf{W}_Y^T(\mathbf{V}_Y/(\mathbf{W}_Y\mathbf{H}_Y)) \tag{19}$$

where $\mathbf{1}$ is a $K \times L_Y$ matrix with all entries equal to one. Note that (18) and (19) appear respectively in the denominator and numerator in (3). Next, we derive the gradient of the regularization term $\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ in (16) with respect to $\mathbf{H}_Y$. Note that by using the equality in (11), i.e., $\mathcal{L}(\bar{\mathbf{V}}|\boldsymbol{\theta}) = \mathcal{L}_B(\bar{\mathbf{V}}|\boldsymbol{\theta})$ for $q(z_i) = \gamma_{il}$, the gradient of $\mathcal{L}(\bar{\mathbf{V}}|\boldsymbol{\theta})$ is identical to that of $\mathcal{L}_B(\bar{\mathbf{V}}|\boldsymbol{\theta})$, which is equivalent to the gradient of $\mathcal{L}_C(\bar{\mathbf{V}}|\boldsymbol{\theta})$. Consequently, the gradient of (16) can be shown in terms of the gradients of $\mathcal{L}_C(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S)$ and $\mathcal{L}_C(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_S)$ with respect to $\mathbf{H}_S$ and $\mathbf{H}_N$, respectively, as,

$$\nabla_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \alpha_S \nabla_{\mathbf{H}_S}\mathcal{L}_C(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S) \\ \alpha_N \nabla_{\mathbf{H}_N}\mathcal{L}_C(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_N) \end{bmatrix} \tag{20}$$

where $\mathcal{L}_C(\cdot|\cdot)$ is the expected LLF given in (13). As we can see from (9), the observations $\bar{\mathbf{V}}_S$ and $\bar{\mathbf{V}}_N$ are expressed in terms of the corresponding basis and activation matrices. Hence, using (13), we can derive the gradients of the expected LLF with respect to the activation matrix in (20), which is shown as

$$\nabla_{\mathbf{H}}\mathcal{L}_C = \nabla^+_{\mathbf{H}}\mathcal{L}_C - \nabla^-_{\mathbf{H}}\mathcal{L}_C \tag{21}$$

where $\mathbf{H}$ stands for either $\mathbf{H}_S$ or $\mathbf{H}_N$, and the dependence of $\mathcal{L}_C(\bar{\mathbf{V}}|\boldsymbol{\theta})$ on $\bar{\mathbf{V}}$ and $\boldsymbol{\theta}$ is omitted for convenience. In (21), the entries of the gradient terms on the right-hand side are

$$[\nabla^+_{\mathbf{H}}\mathcal{L}_C]_{ml} = \sum_{k=1}^{K}\sum_{i=1}^{I}\gamma_{il}\sigma_{i,k}^{-2}\left(\mu_{i,k}\frac{w_{km}}{c_l} + \frac{([\mathbf{WH}]_{kl})^2}{c_l^3}\right) \tag{22}$$

$$[\nabla^-_{\mathbf{H}}\mathcal{L}_C]_{ml} = \sum_{k=1}^{K}\sum_{i=1}^{I}\gamma_{il}\sigma_{i,k}^{-2}(w_{km} + \mu_{i,k})\frac{[\mathbf{WH}]_{kl}}{c_l^2} \tag{23}$$

where $\gamma_{il}$ is the posterior distribution given in (12) and $c_l = \sum_m h_{ml}$ is the normalizing factor. Specifically, $\gamma_{il}$ is computed based on $\mathbf{W}_S$ and $\mathbf{W}_S$ obtained during the training stage and $\mathbf{H}_Y$ estimated in the previous multiplicative update iteration. Note that, based on the concept of the lower bound in (11) and the objective used in the M-step given by (13), the posterior $\gamma_{il}$ is considered as a fixed constant value during the derivations of (22) and (23).[5]

Based on the heuristic multiplicative update rules given in (1), the update rule of $\mathbf{H}_Y$ can be written as,

$$\hat{\mathbf{H}}_Y \leftarrow \hat{\mathbf{H}}_Y \otimes \frac{\nabla^-_{\mathbf{H}_Y}D_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\hat{\mathbf{H}}_Y) + \nabla^+_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \hat{\mathbf{H}}_Y)}{\nabla^+_{\mathbf{H}_Y}D_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\hat{\mathbf{H}}_Y) + \nabla^-_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \hat{\mathbf{H}}_Y)} \tag{24}$$

where $\nabla^+_{\mathbf{H}_Y}D_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y)$ and $\nabla^-_{\mathbf{H}_Y}D_{KL}(\mathbf{V}_Y, \mathbf{W}_Y\mathbf{H}_Y)$ are given in (18) and (19). The components $\nabla^+_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ and $\nabla^-_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y)$ are easily found by substituting (21) into (20). That is,

$$\nabla^+_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \alpha_S \nabla^+_{\mathbf{H}_S}\mathcal{L}_C(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S) \\ \alpha_N \nabla^+_{\mathbf{H}_N}\mathcal{L}_C(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_N) \end{bmatrix} \tag{25}$$

$$\nabla^-_{\mathbf{H}_Y}\mathcal{R}_Y(\mathbf{W}_Y, \mathbf{H}_Y) = \begin{bmatrix} \alpha_S \nabla^-_{\mathbf{H}_S}\mathcal{L}_C(\bar{\mathbf{V}}_S|\boldsymbol{\theta}_S) \\ \alpha_N \nabla^-_{\mathbf{H}_N}\mathcal{L}_C(\bar{\mathbf{V}}_N|\boldsymbol{\theta}_N) \end{bmatrix} \tag{26}$$

where $\nabla^+_{\mathbf{H}_{(\cdot)}}\mathcal{L}_C(\cdot|\cdot)$ in (25) and $\nabla^-_{\mathbf{H}_{(\cdot)}}\mathcal{L}_C(\cdot|\cdot)$ in (26) are given in (22) and (23), respectively.

It is easy to show that the update rule given in (24) takes on non-negative values. In fact, since the posterior distribution and all elements of the mean vector and the diagonal entries of the covariance matrix are non-negative, the values given in (22) and (23) are non-negative. Moreover, the values in (18) and (19) are also non-negative, and therefore the activation matrix is updated under the non-negative elements constraint.

After estimating the activation matrix of the noisy speech, the smoothed PSDs of both the clean speech and noise, $\hat{\mathbf{P}}_S$ and $\hat{\mathbf{P}}_N$, are obtained by using (7) and (8). Then the clean speech spectrum is estimated by Wiener filtering as given in (6). This proposed algorithm based on regularized NMF with Gaussian mixtures will be referred to as RNG.

## 4.2. RNG with weighted Wiener filtering

In this subsection, we describe our second method which uses a WWF. First, the masking threshold estimation is described in Section 4.2.1, and then we introduce the proposed WWF in Section 4.2.2.

### 4.2.1. Masking threshold estimation

The masking effect, which is a psychoacoustical property of the human auditory system, has been employed in diverse applications such as audio and speech coding (Painter and Spanias, 2000) and speech enhancement (Hu and Loizou, 2004; Jabloun and Champagne, 2003; Virag, 1999). Masking refers to a process where one sound is rendered inaudible (maskee) due to the presence of another sound (masker) (Fastl and Zwicker, 2007). The masking properties are modeled using a masking threshold, where the components below the threshold are not perceived. There are two main masking phenomena, simultaneous (spectral) and non-simultaneous (temporal) masking. The former occurs whenever two or more stimuli are simultaneously presented to the auditory system. The latter takes place in the time domain, where the masking occurs both prior and after the onset and offset of the masker with finite duration (Fastl and Zwicker, 2007). In this paper, we only consider the simultaneous masking effect.

Simultaneous masking can be explained in terms of critical band analysis which is a central mechanism in the inner ear. The critical band is specified by means of the so-called Bark scale, which is a perceptual measure relating acoustical frequency to the nonlinear perceptual resolution, in which one Bark covers one critical band. The analytical expression of the mapping function from the frequency $f$ [kHz] to the Bark frequency $b$ [Bark] is shown as

$$b(f) = 13\,\arctan(0.76f) + 3.5\,\arctan[(f/7.5)^2]. \tag{27}$$

[5] Alternatively, we can derive the gradient terms directly from (16), which also lead to (22) and (23).

We followed the procedure introduced in Painter and Spanias (2000) for evaluating the masking threshold in the $l$th time frame, where we here briefly summarize the different steps involved in the computation; further implementation details are given in Painter and Spanias (2000).

(1) *Spectral analysis and normalization*: The PSD is normalized and presented in dB scale as,

$$\bar{P}(k,l) = 90.302 + 10\log_{10}[\hat{P}_S(k,l)/L_w^2] \tag{28}$$

where $L_w$ denotes the analysis window length for the STFT, the constant 90.302 is used for the power compensation, and $\hat{P}_S(k,l)$ is the estimated clean speech PSD given in (7).

(2) *Identification of tonal and non-tonal maskers*: Tonal maskers are identified according to the local maxima of the normalized PSD, $\bar{P}(k,l)$. A single non-tonal (noise-like) masker for each critical band is then identified by summing the energy of the spectral components which have not contributed to a tonal masker.

(3) *Reorganization of maskers*: Any tonal or non-tonal maskers below the absolute hearing threshold (AHTH) are discarded, where the AHTH in dB versus frequency $f$ [kHz] is shown as

$$T_A(f) = 3.65 f^{-0.8} - 6.5 e^{-0.6(f-3.3)^2} + 10^{-3} f^4 \tag{29}$$

Next, any pair of maskers within a distance of 0.5 Bark are replaced by the stronger of the two.

(4) *Individual masking threshold*: The individual masking threshold at frequency bin $i$ due to a tonal masker at frequency bin $j$ is given in dB as

$$T_{tm}(i,j) = \bar{P}_{tm}(j) - 0.275\ b(f_j) + \mathrm{SF}(i,j) - 6.025 \tag{30}$$

where $\bar{P}_{tm}(j)$ is the level of tonal masker, $f_j$ [kHz] is the corresponding frequency of the $j$th bin, $b(f_j)$ denotes the Bark frequency given in (27) and SF($i$, $j$) is the spreading function which accounts for the inter-band masking. The latter is given as

$$\mathrm{SF}(i,j) = \begin{cases} 17\Delta_b - 0.4\bar{P}_{tm}(j) + 11, & -3 \le \Delta_b < -1 \\ (0.4\bar{P}_{tm}(j) + 6)\Delta_b, & -1 \le \Delta_b < 0 \\ -17\Delta_b, & 0 \le \Delta_b < 1 \\ (0.15\bar{P}_{tm}(j) - 17)\Delta_b - 0.15\bar{P}_{tm}(j), & 1 \le \Delta_b < 8 \end{cases} \tag{31}$$

where $\Delta_b = b(f_i) - b(f_j)$. Similarly, the masking threshold of a non-tonal masker is given as,

$$T_{nm}(i,j) = \bar{P}_{nm}(j) - 0.175\ b(f_j) + \mathrm{SF}(i,j) - 2.025 \tag{32}$$

where $\bar{P}_{nm}(j)$ is the non-tonal masker level. The spreading function used in (32) is identical to (31) where $\bar{P}_{tm}(j)$ is replaced by $\bar{P}_{nm}(j)$. The above computation of the masking thresholds $T_{tm}(i,j)$ for tonal maskers and $T_{nm}(i,j)$ for non-tonal ones are repeated for each frame; whenever such a computed threshold value falls below the AHTH, it is replaced by the latter.

(5) *Global masking threshold*: Finally, the resulting individual masking thresholds are summed linearly along with the AHTH to obtain the global masking threshold in dB in the $k$th frequency bin, which is shown as,

$$T_g(k,l) = 10\log_{10}\left(10^{0.1T_A(f_k)} + \sum_{n=1}^{N_{tm}} 10^{0.1T_{tm}(k,j_n)} + \sum_{n=1}^{N_{nm}} 10^{0.1T_{nm}(k,j_n)}\right) \tag{33}$$

where $N_{tm}$ and $N_{nm}$ respectively denote the number of tonal and non-tonal maskers and $j_n$ is the frequency bin location of the $n$th masker. An example of the global masking threshold is illustrated in Fig. 1, where we considered a speech signal of a female speaker.

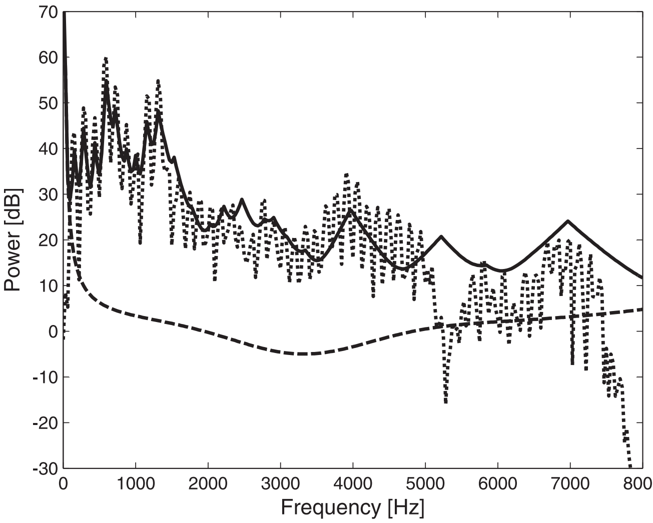

**Fig. 1.** Example of masking threshold (dotted: normalized power spectrum of a female speaker, solid: masking threshold, dashed: absolute hearing threshold).

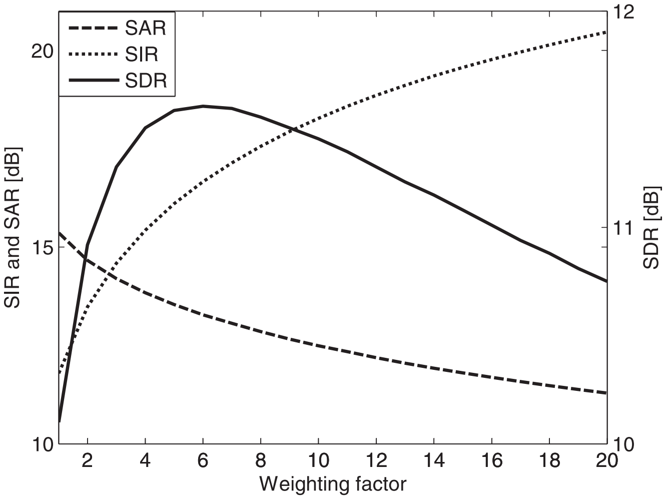

**Fig. 2.** SDR, SIR and SAR values for different weighting factors in WWF.

### 4.2.2. Weighted Wiener filtering

A generalized Wiener filtering has been introduced in Lim and Oppenheim (1979), which is shown as,

$$\hat{S}(k,l) = \left(\frac{\hat{P}_S(k,l)}{\hat{P}_S(k,l) + \eta\hat{P}_N(k,l)}\right)^{\nu} Y(k,l) \tag{34}$$

where $\eta$ and $\nu$ are tuning parameters. For simplicity, we will fix $\nu$ to 1 in the proposed framework, and refer to the resulting method as weighted Wiener filtering (WWF) (Spriet et al., 2005). The weighting factor $\eta$ is known to control the trade-off between noise reduction and speech distortion. For a large $\eta$, for instance, more noise reduction is performed at the expense of increased speech distortion, and vice versa. This phenomenon is illustrated in Fig. 2 where we computed different objective measures while varying $\eta$ from 1 to 20. The objective measures considered are the source-to-interference ratio (SIR), source-to-artifact ratio (SAR) and source-to-distortion ratio (SDR) (Vincent et al., 2006).[6] The noisy speech was generated by adding a factory noise to selected clean speech files[7] at a 5 dB input SNR, and the results were obtained by averaging over different speakers. For each noisy speech, the clean speech and noise PSDs were computed from the proposed RNG method introduced in Section 4.1, followed by temporal smoothing given in (7) and (8). As we can see from Fig. 2, the results obtained for the different objective measures vary greatly as a function of $\eta$ and therefore, an appropriate selection of the weighting factor is necessary.

In contrast to using a constant value as the weighting factor in (34), it has been proposed to select different weighting factor for each time-frequency bin, i.e., $\eta(k, l)$, based on the masking threshold computed for each of these bins. Gustafsson et al. (1998) proposed a heuristic approach where the linear estimator of the clean speech spectrum was derived, aiming to mask the distortion of the residual noise which is defined as the difference between the actual and residual noise powers. This estimator was extended in Hu and Loizou (2004) by solving an optimization problem which minimizes a related error criterion. Defraene et al. (2012) proposed to use an exponential function to map the so-called noise-to-mask ratio (NMR) into the weighting factor, where the NMR in dB, $\Phi(k, l)$, is defined as the log distance from the minimum masking threshold in one critical band to the noise level (Painter and Spanias, 2000):

$$\Phi(k,l) = \bar{P}_N(k,l) - \min_{k\in C_b} T_g(k,l) \tag{35}$$

where $C_b$ is the set of frequency bins for the $b$th critical band and $\bar{P}_N(k,l)$ is the normalized PSD given in (28).

For all these algorithms, a zero weighting factor is applied when the noise power is lower than the masking threshold, i.e., $\eta(k,l) = 0$ for $T_g(k,l) > \bar{P}_N(k,l)$. However, this strict condition limits the performance, since the masking threshold is calculated from an inaccurate estimate of the clean speech PSD. Although we can expect that a more accurate clean speech PSD can be obtained by using the proposed RNG method, we further suggest to relax this strict condition by taking into account in a continuous way the case where the noise power is even lower than the masking threshold. This approach can be regarded as a *soft* decision on the weighting factor.

In advance of describing the proposed method, we summarize several intuitive aspects, which should be considered for selecting the weighting factors in the WWF, as follows. When $T_g(k, l)$ is low, the noise signal (maskee) is easily perceived due to the low masking capability of the speech signal (masker). The emphasis then should be put on reducing this perceivable noise. Consequently, a high weighting factor is necessary in the WWF. On the contrary, if $T_g(k, l)$ is high, the noise is easily masked by the speech. Hence, a small weighting factor is selected. Note that these aspects hold for both the cases where the NMR is either positive or negative. The difference is that a much smaller weighting factor for the case of negative NMR is necessary compared to the positive NMR.

In the proposed WWF, the weighting factor is selected through a heuristic approach using a sigmoid function as a mapping from the NMR to the weighting factor. The motivation for using the sigmoid function is to limit the range of the weighting factor to be selected, therefore avoiding extreme values that could lead to instability (Fig. 3). The proposed mapping function is given by

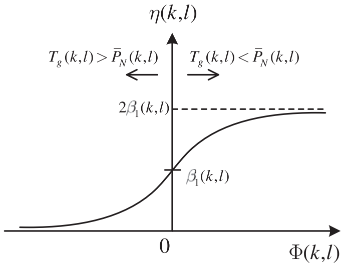

**Fig. 3.** Proposed mapping function from NMR, $\Phi(k, l)$, to weighting factor, $\eta(k, l)$, based on a sigmoid function.

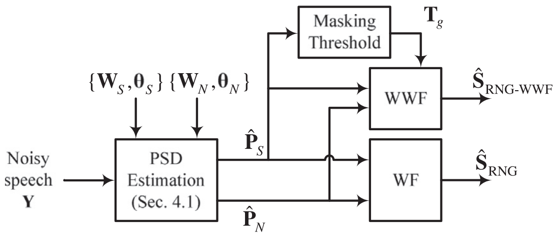
Masking Threshold
$\mathbf{T}_g$
$\{\mathbf{W}_S, \boldsymbol{\theta}_S\}$ $\{\mathbf{W}_N, \boldsymbol{\theta}_N\}$
WWF
$\hat{\mathbf{S}}_{\text{RNG-WWF}}$
Noisy speech $\mathbf{Y}$
PSD Estimation (Sec. 4.1)
$\hat{\mathbf{P}}_S$
$\hat{\mathbf{P}}_N$
WF
$\hat{\mathbf{S}}_{\text{RNG}}$

**Fig. 4.** Simplified block diagrams of RNG and RNG-WWF methods.

$$\eta(k,l) = \frac{2\beta_1(k,l)}{1+\exp(-\beta_2(k,l)\Phi(k,l))} \tag{36}$$

where $\beta_1(k, l)$, $\beta_2(k, l) > 0$ are tuning parameters and the NMR, $\Phi(k, l)$, is given in (35). The value $\beta_1(k, l)$ defines the range of $\eta(k, l) \in (0, 2\beta_1(k, l))$ and $\beta_2(k, l)$ determines the slope of the sigmoid function. For simplicity of the implementation, we consider a constant slope, i.e., $\beta_2(k,l) = \beta_2$, and identical values of $\beta_1(k, l)$ across the frequency bins for a given time frame, i.e., $\beta_1(k,l) = \beta_1(l)$.

The value $\beta_1(l)$ is calculated using the following function,

$$\beta_1(l) = \rho_1 e^{-\rho_2 R(l)} \tag{37}$$

where $\rho_1$, $\rho_2 > 0$ are tuning parameters and $R(l)$ is defined as

$$R(l) = 10\log_{10}\frac{\sum_k \hat{P}_S(k,l)}{\sum_k \hat{P}_N(k,l)}. \tag{38}$$

The underlying motivation for using the form given in (37) and (38) is similar to the approach introduced in Kodrasi et al. (2015). That is, a small weighting factor is selected for a high input SNR. Specifically in the proposed method, the input SNR for a given time frame of the noisy speech is estimated from $R(l)$ given in (38), which is then applied to determine the range of $\eta(k, l)$ through $\beta_1(l)$ given in (37).

The proposed enhancement algorithm based on the regularized NMF with Gaussian mixtures and weighted Wiener filtering will be referred to as RNG-WWF. A simplified block diagram of both the RNG and RNG-WWF methods is illustrated in Fig. 4. We note that for both algorithms, the same training approach as described in Section 3 is employed.

## 5. Experiments

In this section, a performance evaluation of the proposed methods is presented.

### 5.1. Methodology

We used clean speech from the TSP (Kabal, 2002) and Grid Corpus (Cooke et al., 2006) databases and noise from the NOISEX database (Varga and Steeneken, 1993), where the sampling

[6] For a given target source, the interference refers to unwanted signal components such as noise, whereas the artifact refers to components caused by other phenomena, such as e.g., forbidden distortion. In speech enhancement applications, these measures can be interpreted as follows: the SIR and SAR are proportional to the amount of noise reduction and inversely proportional to the speech distortion, respectively, while SDR measures the overall quality of the enhanced speech (Mohammadiha et al., 2013)

[7] Further details about various speech and noise files used in our experimental work are described in detail in Section 5.

**Table 1**
A comparison between different perceptually-motivated and/or weighting methods.

| Reference | Gain function, $G(k, l)$<br>$(\hat{S}(k,l) = G(k,l)Y(k,l))$ | Description |
|---|---|---|
| Gustafsson et al. (1998) | $\min\left(\sqrt{\frac{T_g(k,l)}{\hat{P}_N(k,l)}} + \zeta, 1\right)$ | Heuristic gain function, aiming to mask the distortion of the residual noise |
| Hu and Loizou (2004) | $\left(1 + \max\left(\sqrt{\frac{\hat{P}_N(k,l)}{T_g(k,l)}} - 1, 0\right)\right)^{-1}$ | Gain function obtained by minimizing an error criterion (extension of Gustafsson et al., 1998) |
| Defraene et al. (2012) | $\frac{\hat{P}_S(k,l)}{\hat{P}_S(k,l) + \eta(k,l)\hat{P}_N(k,l)}$ | Heuristic mapping from the NMR to $\eta(k, l)$ (*hard* decision) |
| Kodrasi et al. (2015) | | Curvature-based optimization for the estimation of $\eta(k, l)$ |
| Proposed | | Heuristic mapping from the NMR to $\eta(k, l)$ (*soft* decision) |

rate of all signals was adjusted to 16 kHz. For the clean speech, 20 speakers (10 males and 10 females) were selected from the TSP and 34 speakers (17 males and 17 females) from the Grid Corpus databases for a total of 54 speakers. For the noises, we selected the buccaneer 1, hfchannel, babble and factory 1 noises from the NOISEX database. Each clean speech and noise signal was divided into three disjoint groups: i) *training data*, used for estimating the NMF and GMM parameters, ii) *validation data*, used for selecting the regularization coefficients and tuning parameters, and iii) *test data*, used for final verification. Specifically, the training data consisted of approximately 2 min (50 sentences) and 8 min (350 utterances) of long speech segments for each speaker from the TSP and Grid Corpus databases, respectively, as well as 3 min segment for the noises. The validation data consisted of 12 s (5 sentences) and 20 s (15 utterances) of speech for each speaker from the TSP and Grid Corpus databases, respectively, and 30 s of noise from the NOISEX database. The same partitioning was used for the test data. The noisy speech signals were generated from the test and validation signals by scaling and adding the noise to the clean speech (based on the estimated variances of the time-domain signals) to obtain input SNRs of 0, 5 and 10 dB. The STFT analysis was implemented by using a Hanning window of 512 samples with 50 % overlap. After enhancement, the estimated clean speech signal in the time-domain was reconstructed by applying the inverse STFT on its spectrum followed by the overlap-add method.

Regarding the implementation of the proposed algorithms, we considered a speaker-dependent (SD) application, where one basis matrix and associated GMM parameter set were trained for each speaker. We used $M = 80$ basis vectors and $I = 8$ Gaussian components in the GMM for both the clean speech and noise. The values of $(\tau_S, \tau_N) = (0.4, 0.9)$ were chosen empirically using the validation set and used as the temporal smoothing factors in (7) and (8). For the regularization coefficients $\alpha_S$ and $\alpha_N$ in (16), we examined different values from 0.0005 to 0.1 and obtained good results in the range [0.005, 0.01]. Hence, we selected $(\alpha_S, \alpha_N) = (0.005, 0.01)$. We also examined several choices for the tuning parameters in the proposed weighting function (36), i.e. $\rho_1$, $\rho_2$ and $\beta_2$. We first fixed $\rho_1$ to 4, 5 and 6, based on the results shown in Fig. 2. For each value of $\rho_1$, we then considered various choices of $\beta_1$ and $\rho_2$ and determined the ones that gave the highest SDR values. Good results for both $\beta_2$ and $\rho_2$ were found around [0.005, 0.1]. Ultimately, we chose $\beta_2 = 0.01$ and $(\rho_1, \rho_2) = (5, 0.1)$ for the experiments.

We used the PESQ (Recommendation, 2001), SDR (Vincent et al., 2006), as well as the segmental SNR as the objective measures of performance. The PESQ attempts to predict overall perceptual quality in mean opinion score (MOS) and the SDR measures the overall quality of the enhanced speech in dB by considering both the speech distortion and noise reduction as explained in Section 4.2.2. For all the measures, a higher value indicates a better result.

### 5.2. Benchmark algorithms

To evaluate the speech enhancement performance of the newly proposed algorithms, we compared them against several algorithms from the literature. Basic settings such as the STFT analysis and synthesis, number of basis vectors and Gaussian components in the GMM, and masking threshold calculations, when applicable, were kept identical for all the benchmark and proposed algorithms. Also, we considered the SD application for all NMF-based algorithms.

The benchmark algorithms were categorized into two groups. The purpose of the first group was only to compare the enhancement performance of the proposed WWF (i.e., RNG-WWF) to that of other perceptually-motivated and/or weighting methods. Specifically, we considered the algorithms proposed by Gustafsson et al. (1998), Hu and Loizou (2004), Defraene et al. (2012), Kodrasi et al. (2015); in the sequel, we shall refer to each algorithm using the names of its authors for simplicity. Although the algorithms in Defraene et al. (2012) and Kodrasi et al. (2015) were proposed for multi-channel speech enhancement, they can still be applied in the current single-channel framework. We used the following tuning parameters for these algorithms: a trade-off control parameter $\zeta = 0.1$ in Gustafsson et al. (1998), $(\gamma, \delta, \epsilon) = (0.2, 0.9, 0.9)$ in Defraene et al. (2012) and $(\alpha, \beta) = (1, 2)$ in Kodrasi et al. (2015) (see the references for the meaning of these notations). For all the benchmark algorithms and RNG-WWF method, we employed identical PSDs of the clean speech and noise, which were estimated using the RNG method. The salient features of the benchmarks and proposed algorithms are summarized in Table 1.

The purpose of the second group was to compare the enhancement performance of the proposed algorithms with that of various speech enhancement algorithms, which are given below. Note that, for all NMF-based algorithms, except the proposed RNG-WWF method which requires a weighting factor, we used the same reconstruction method introduced in Section 2, i.e., computing smoothed PSDs and Wiener filtering, for fair comparison.

(1) *Short-time spectral amplitude estimator (STSA)*: We implemented the well-known classical STSA estimator proposed by Ephraim and Malah (1984). A smoothing factor of 0.98 in the decision-directed (DD) method for *a priori* SNR estimation was used. The noise PSD was estimated using an algorithm described in Gerkmann and Hendriks (2012) with a value of 0.8 for the smoothing factor.

(2) *Spectral subtraction with masking properties (SSM)*: We considered a spectral subtraction algorithm with masking properties proposed in Virag (1999). The noise PSD in this approach was also estimated using the algorithm from Gerkmann and Hendriks (2012) with 0.8 for the smoothing factor.

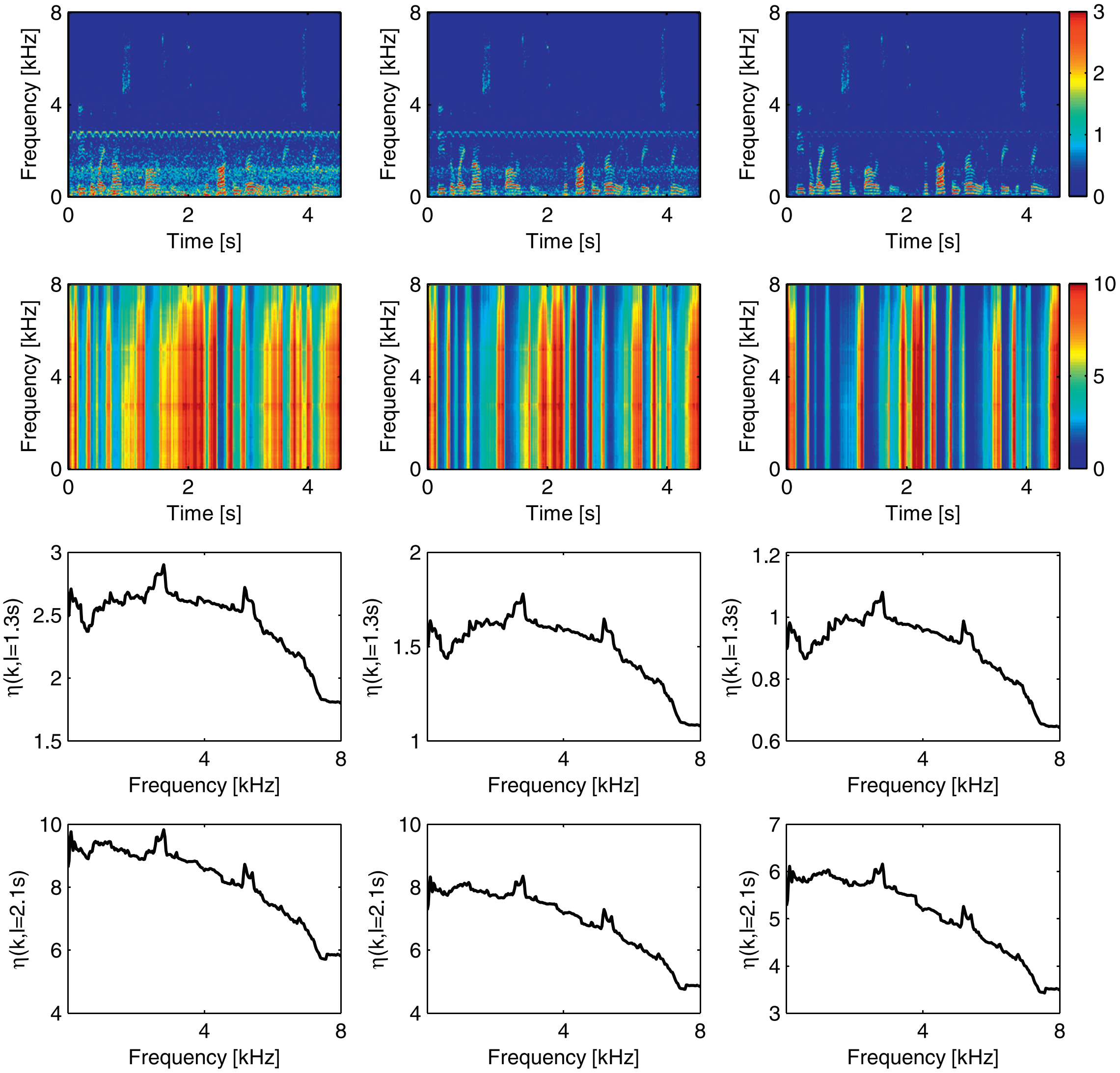

**Fig. 5.** Examples of proposed weighting factor. Each column from left to right respectively correspond to input SNR of 0, 5 and 10 dB. Each row from top to bottom shows the noisy speech magnitude spectrum, time-frequency representation of the proposed weighting factor and the weighting factor at the time frame of 1.3 s and 2.1 s.

(3) *Standard NMF*: The standard NMF algorithm based on KL-divergence introduced in Section 2 was evaluated, which will be referred to as NMF.

(4) *Regularized NMF*: In order to compare with other regularization-based NMF algorithms, we chose an algorithm proposed by Grais and Erdogan (2013), where the column vectors of the activation matrix of the clean speech and noise are modeled by distinct GMMs. We employed the sequential form of training, and used the regularization coefficients of $(\alpha_S, \alpha_N) = (0.005, 0.001)$ in our experiments as they provided good results. This method will be referred to as RNMF-AGM.

(5) *Weighted NMF (WNMF)*: We evaluated a perceptually weighted NMF (WNMF) algorithm introduced in Virtanen (2007b), where the perceptual weighting matrix was constructed (based on the masking threshold) as in Nikunen and Virtanen (2010). Although the WNMF algorithm was originally proposed for an unsupervised application, we applied it in a supervised manner. That is, the basis matrices for the clean speech and noise were obtained independently during the training stage. In the enhancement stage, the WNMF activation update was applied to the noisy speech, where the masking threshold was calculated from the noisy speech. Although the masking threshold can be obtained from the estimated clean speech PSD by first applying a simple speech enhancement scheme (Defraene et al., 2012; Virag, 1999), we followed the original paper, since we observed similar results when using the masking threshold either computed from the noisy or estimated clean speech PSD.

### 5.3. Results

We first illustrate an example of the proposed weighting factor $\eta(k, l)$ for different input SNRs in Fig. 5. In this particular example, a male speech is degraded with buccaneer 1 noise at 0, 5 and 10 dB input SNR. We can make the following observations:

- The values of $\eta(k, l)$ around 3 kHz, which corresponds to the intense ringing sound of the buccaneer 1 noise, are larger compared to the other frequencies;
- For a given time-frequency bin, $\eta(k, l)$ decreases as the input SNR increases from 0 to 10 dB;
- The values of $\eta(k, l)$ at the time frame of 2.1 s (a speech-absence period) are larger than the ones at 1.3 s (a speech-presence period).

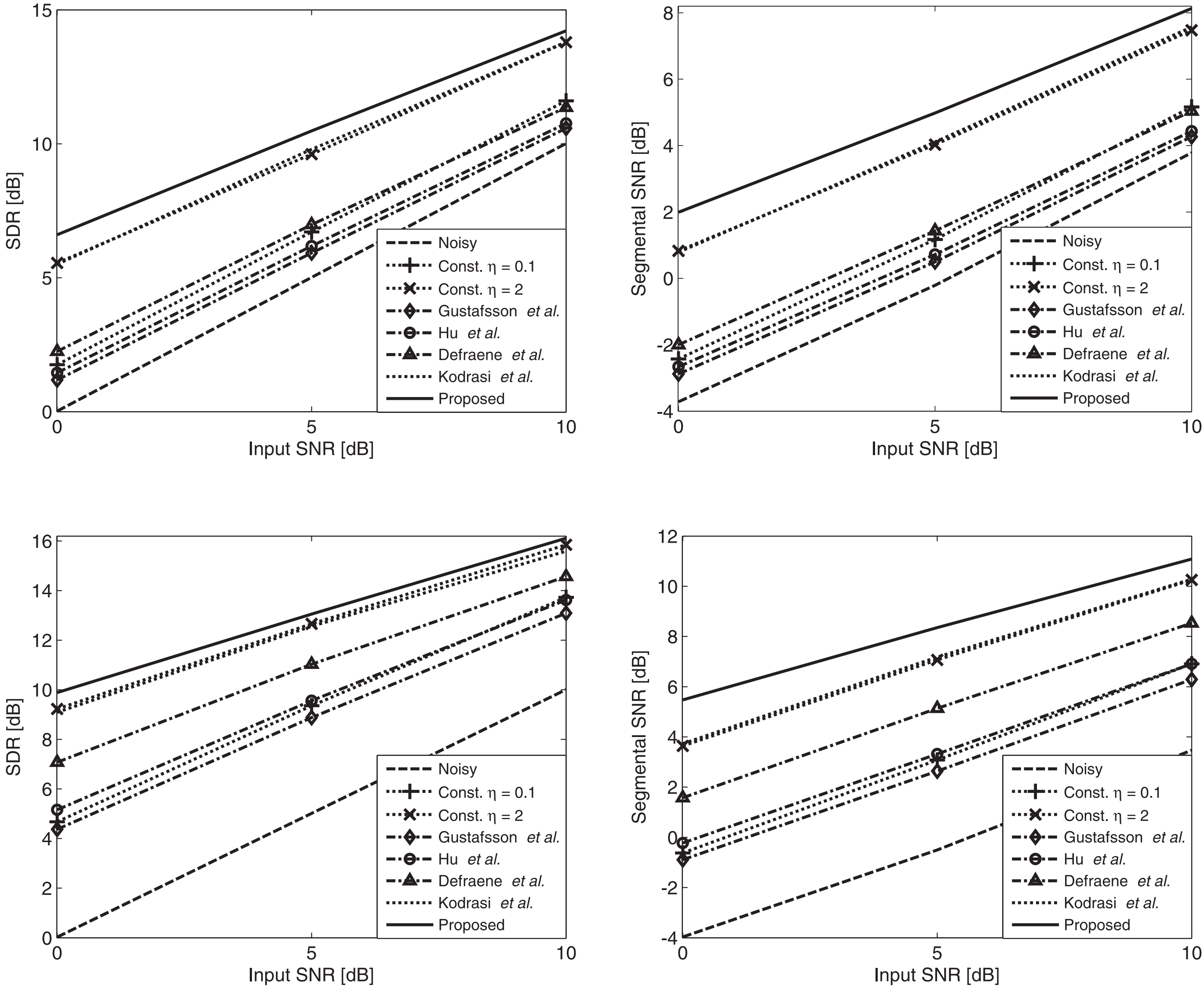

**Fig. 6.** SDR and segmental SNR comparisons for factory 1 (top) and hfchannel (bottom) noises.

These phenomena are essentially due to the estimated input SNR $R(l)$ given by (38). That is, as we intended, a larger value of $\eta(k, l)$ is selected based on (36) and (37), for a lower value of $R(l)$. Consequently, the noise components will be further suppressed in the corresponding time-frequency bins.

We compared the proposed RNG-WWF method with other methods in the first group of benchmark algorithms in order to verify the performance of the proposed weighting method. Average SDR and segmental SNR values over all speakers for factory 1 and hfchannel noises, with 0, 5 and 10 dB input SNRs, are displayed in Fig. 6. We can see that in all cases, the proposed weighting scheme provides the best results. It is worth noting that the perceptually-motivated benchmark algorithms showed a worse performance than using a constant weighting factor of $\eta = 2$, and tend to show similar quality to using $\eta = 0.1$. This is mainly due to the hard decision on the weighting factor such that $\eta(k, l) = 0$ for $\bar{P}_N(k,l) < T_g(k,l)$, which leads to $\hat{S}(k,l) = Y(k,l)$, i.e., the noise components are not reduced in such time-frequency bins. Therefore, it is verified through experiments that employing soft decision on the weighting factor, i.e., applying non-zero value on $\eta(k, l)$ for $\bar{P}_N(k,l) < T_g(k,l)$, improves the enhancement performance. Similar results were also found for the babble and buccaneer 1 noises.

Regarding the benchmark algorithms in the second group and the proposed algorithms, the average results over all speakers of the three objective measures (i.e., PESQ, SDR and segmental SNR) are shown for each noise type, respectively, in Tables 2–5. The values in bold indicate the best performance along the row. As it can be observed, the best enhancement results were obtained with the proposed RNG-WWF method for all the different noise types and input SNRs. Moreover, the RNG method generally provided better results than the benchmark algorithms except in specific cases, e.g., segmental SNR for the factory 1 noise at 0 dB input SNR. Among the benchmark algorithms, the STSA and SSM which used no training data provided reasonable results for babble and factory 1 noises compared to the NMF-based algorithms. However, they resulted in poorer performances for buccaneer 1 and hfchannel noises. Among the NMF-based benchmark algorithms, which used training data to obtain some prior knowledge of the clean speech and noise, it was found in general that the RNMF-AGM provided slightly better results compared to the NMF and WNMF methods (except in some cases, e.g., slightly better PESQ results using the WNMF method for the buccaneer 1 and factory 1 noises). If we only compare between the two proposed methods, the RNG-WWF method provided much better results than the RNG method,

**Table 2**
Average results for buccaneer 1 noise.

| Input<br>SNR | Eval. | Noisy | STSA | SSM | NMF | WNMF | RNMF<br>-AGM | RNG | RNG<br>-WWF |
|---|---|---|---|---|---|---|---|---|---|
| 0 dB | PESQ | 1.25 | 1.58 | 1.61 | 1.79 | 1.83 | 1.81 | 1.98 | **2.22** |
| | SDR | 0.02 | 4.31 | 4.25 | 5.25 | 5.74 | 5.43 | 6.13 | **7.92** |
| | SNRseg | −3.97 | −0.27 | −0.56 | 0.13 | 1.15 | 0.28 | 1.79 | **3.18** |
| 5 dB | PESQ | 1.54 | 1.94 | 1.99 | 2.18 | 2.21 | 2.20 | 2.35 | **2.47** |
| | SDR | 5.01 | 8.56 | 8.79 | 9.75 | 9.63 | 9.92 | 10.59 | **11.38** |
| | SNRseg | −0.49 | 2.79 | 2.78 | 3.58 | 4.07 | 3.75 | 4.40 | **6.17** |
| 10 dB | PESQ | 1.89 | 2.32 | 2.39 | 2.53 | 2.55 | 2.55 | 2.64 | **2.69** |
| | SDR | 10.01 | 12.43 | 12.97 | 13.80 | 13.23 | 13.91 | 14.59 | **14.85** |
| | SNRseg | 3.48 | 6.14 | 6.47 | 7.14 | 7.28 | 7.33 | 8.06 | **9.19** |

**Table 3**
Average results for hfchannel noise.

| Input<br>SNR | Eval. | Noisy | STSA | SSM | NMF | WNMF | RNMF<br>-AGM | RNG | RNG<br>-WWF |
|---|---|---|---|---|---|---|---|---|---|
| 0 dB | PESQ | 1.23 | 1.50 | 1.59 | 1.78 | 1.71 | 1.79 | 2.01 | **2.30** |
| | SDR | 0.03 | 7.11 | 7.62 | 7.32 | 6.97 | 7.51 | 8.31 | **9.88** |
| | SNRseg | −3.97 | 1.95 | 2.35 | 1.64 | 2.16 | 1.81 | 2.56 | **5.46** |
| 5 dB | PESQ | 1.45 | 1.92 | 2.04 | 2.15 | 2.08 | 2.16 | 2.35 | **2.51** |
| | SDR | 5.02 | 10.80 | 11.66 | 11.50 | 10.85 | 11.66 | 12.37 | **13.05** |
| | SNRseg | −0.50 | 4.96 | 5.78 | 5.12 | 5.22 | 5.30 | 6.20 | **8.35** |
| 10 dB | PESQ | 1.75 | 2.31 | 2.46 | 2.50 | 2.43 | 2.52 | 2.63 | **2.70** |
| | SDR | 10.01 | 14.12 | 15.19 | 15.12 | 14.44 | 15.22 | 15.91 | **16.11** |
| | SNRseg | 3.47 | 7.91 | 9.03 | 8.58 | 8.48 | 8.74 | 9.67 | **11.09** |

**Table 4**
Average results for babble noise.

| Input<br>SNR | Eval. | Noisy | STSA | SSM | NMF | WNMF | RNMF<br>-AGM | RNG | RNG<br>-WWF |
|---|---|---|---|---|---|---|---|---|---|
| 0 dB | PESQ | 1.52 | 1.68 | 1.62 | 1.77 | 1.72 | 1.78 | 1.81 | **1.84** |
| | SDR | 0.02 | 2.76 | 2.69 | 3.06 | 2.52 | 3.18 | 3.36 | **4.55** |
| | SNRseg | −3.48 | −0.57 | −0.65 | −0.36 | −0.34 | −0.32 | −0.29 | **1.28** |
| 5 dB | PESQ | 1.86 | 2.05 | 2.02 | 2.16 | 2.11 | 2.17 | 2.20 | **2.24** |
| | SDR | 5.01 | 7.39 | 7.53 | 7.70 | 6.80 | 7.89 | 8.12 | **8.53** |
| | SNRseg | 0.05 | 2.44 | 2.58 | 2.79 | 2.54 | 2.94 | 3.09 | **4.06** |
| 10 dB | PESQ | 2.22 | 2.42 | 2.43 | 2.53 | 2.47 | 2.55 | 2.56 | **2.59** |
| | SDR | 10.01 | 11.52 | 11.90 | 11.53 | 10.38 | 11.73 | 12.17 | **12.21** |
| | SNRseg | 4.05 | 5.84 | 6.23 | 5.91 | 5.66 | 6.16 | 6.66 | **7.07** |

**Table 5**
Average results for factory 1 noise.

| Input<br>SNR | Eval. | Noisy | STSA | SSM | NMF | WNMF | RNMF<br>-AGM | RNG | RNG<br>-WWF |
|---|---|---|---|---|---|---|---|---|---|
| 0 dB | PESQ | 1.36 | 1.68 | 1.66 | 1.74 | 1.80 | 1.76 | 1.80 | **1.98** |
| | SDR | 0.02 | 4.44 | 4.16 | 4.34 | 4.29 | 4.54 | 4.49 | **6.60** |
| | SNRseg | −3.72 | 0.28 | 0.17 | −0.14 | 0.28 | 0.12 | −0.10 | **1.99** |
| 5 dB | PESQ | 1.70 | 2.09 | 2.10 | 2.15 | 2.18 | 2.16 | 2.19 | **2.34** |
| | SDR | 5.01 | 8.62 | 8.69 | 9.07 | 8.53 | 9.24 | 9.27 | **10.48** |
| | SNRseg | −0.21 | 3.21 | 3.34 | 3.33 | 3.19 | 3.53 | 3.42 | **4.99** |
| 10 dB | PESQ | 2.07 | 2.45 | 2.50 | 2.53 | 2.52 | 2.54 | 2.54 | **2.64** |
| | SDR | 10.01 | 12.49 | 12.91 | 13.33 | 12.42 | 13.37 | 13.61 | **14.22** |
| | SNRseg | 3.78 | 6.48 | 6.91 | 6.91 | 6.46 | 6.96 | 7.12 | **8.13** |

which further validates that using the proposed weighting factor improves the enhanced speech quality.

Fig. 7 illustrates the magnitude spectra of clean, noisy and enhanced speech for several benchmark and proposed algorithms. In this particular example, a female speech is degraded with buccaneer 1 noise at 0 dB input SNR. As we can see, the proposed RNG-WWF method could reduce the background noise significantly, and especially during the speech-absence periods where the noise is further reduced.

Informal listening tests were also conducted to compare the performance of the benchmark algorithms in the second group and the proposed algorithms. It was generally found that the latter, and especially the RNG-WWF method offered the best performance, both in terms of noise reduction and speech distortion. More specifically, the STSA and SSM gave an enhanced speech with reasonable quality for the babble and factory 1 noises although some musical noise was found in the SSM method. However, they both failed to remove high frequency components in the buccaneer 1 noise which resulted in a highly annoying ringing sound. The

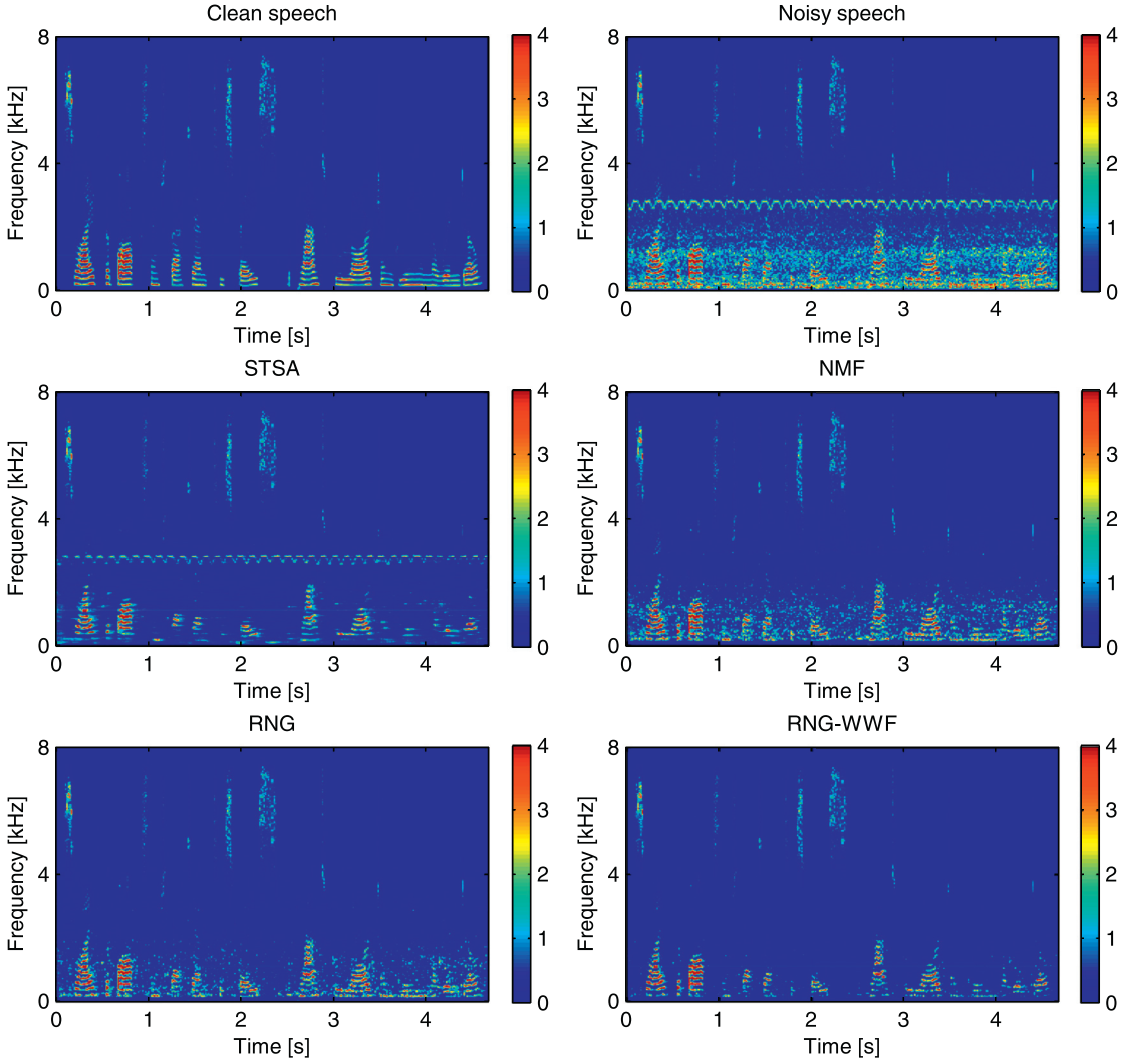

**Fig. 7.** Example of magnitude spectra of the clean, noisy and estimated clean speech for the benchmark and proposed algorithms. A female speech is degraded with buccaneer 1 noise at 0 dB input SNR.

enhanced speech with the benchmark NMF algorithms, i.e., NMF, RNMF-AGM and WNMF, was perceived as being similar to that obtained with the STSA and SSM for babble and factory 1 noises, but of better quality for buccaneer 1 and hfchannel noises. Focusing on the proposed algorithms, the RNG method could remove more low frequency noise than the benchmark algorithms, whereas the high frequency components were further removed using the RNG-WWF method. Consequently, the enhanced speech using the RNG-WWF method was perceived as having much better quality than the one using the RNG method.

## 6. Conclusion

New single-channel speech enhancement algorithms based on regularized NMF have been introduced. In the proposed framework, *a priori* knowledge about the magnitude spectra of the clean speech and noise is captured by distinct GMMs, where normalized spectra are employed to handle the magnitude difference between the training and test data. The corresponding LLFs are included as regularization terms in the NMF cost function during the enhancement stage. Further improvement of the enhanced speech quality was obtained by exploiting the masking effects of the human auditory system. Specifically, we constructed a weighted Wiener filter where the weighting factor is selected based on the masking threshold calculated from estimated clean speech PSD. In addition to informal listening tests and visual inspection of spectrograms, experimental results using three different objective measures (PESQ, SDR, and segmental SNR) showed that the proposed speech enhancement algorithms could provide better performance than the benchmark algorithms for several types of noises and input SNRs.

## References

Bertin, N., Badeau, R., Vincent, E., 2010. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. IEEE Trans. Audio Speech Lang. Process. 18 (3), 538–549.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27 (2), 113–120.

Cemgil, A.T., 2009. Bayesian inference for nonnegative matrix factorisation models. Comput. Intell. Neurosci, no. 4, Article ID 785152, pp. 1–17.

Chung, H., Plourde, E., Champagne, B., 2014. Regularized NMF-based speech enhancement with spectral components modeled by Gaussian mixtures. In: IEEE International Workshop on Machine Learning for Signal Processing, Reims, France, pp. 1–6.

Cichocki, A., Zdunek, R., Amari, S.-i., 2006. New algorithms for non-negative matrix factorization in applications to blind source separation. In: IEEE International Conference on Acoustics Speech and Signal Process, Toulouse, France, pp. 621–624.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11 (5), 466–475.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (5), 2421–2424.

Defraene, B., Ngo, K., van Waterschoot, T., Diehl, M., Moonen, M., 2012. A psychoacoustically motivated speech distortion weighted multi-channel Wiener filter for noise reduction. In: IEEE International Conference on Acoustics Speech and Signal Process, Kyoto, Japan, pp. 4637–4640.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–38.

Ding, G.-H., Wang, X., Cao, Y., Ding, F., Tang, Y., 2005. Speech enhancement based on speech spectral complex Gaussian mixture model. In: IEEE Acoustics Speech and Signal Process, Pennsylvania, USA, pp. 165–168.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266.

Erkelens, J., Jensen, J., Heusdens, R., 2007. Speech enhancement based on Rayleigh mixture modeling of speech spectral amplitude distributions. In: European Signal Processing Conference, Poznan, Poland, pp. 9–65.

Fastl, H., Zwicker, E., 2007. Psychoacoustics: Facts and Models, vol. 22. Springer Science & Business Media.

Févotte, C., Bertin, N., Durrieu, J.-L., 2009. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Comput. 21 (3), 793–830.

FitzGerald, D., Cranitch, M., Coyle, E., 2008. On the use of the Beta divergence for musical source separation. In: Irish Signals and Systems Conference, Galway, Ireland.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. 20 (4), 1383–1393.

Grais, E.M., Erdoğan, H., 2012. Hidden Markov models as priors for regularized nonnegative matrix factorization in single-channel source separation. In: Annual Conference of the International Speech Communication Association. ISCA, Portland, USA, pp. 1536–1539.

Grais, E.M., Erdogan, H., 2013. Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation. Comput. Speech Lang. 27 (3), 746–762.

Gustafsson, S., Jax, P., Vary, P., 1998. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In: IEEE International Conference on Acoustics Speech and Signal Process, Washington, USA, pp. 397–400.

Hansen, J.H., Radhakrishnan, V., Arehart, K.H., 2006. Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. IEEE Trans. Audio Speech Lang. Process. 14 (6), 2049–2063.

Hao, J., Lee, T.-W., Sejnowski, T.J., 2010. Speech enhancement using Gaussian scale mixture models. IEEE Trans. Audio Speech Lang. Process. 18 (6), 1127–1136.

Hermus, K., Wambacq, P., Hamme, H.V., 2007. A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Appl. Signal Process. 2007 (1), 195.

Hu, Y., Loizou, P.C., 2004. Incorporating a psychoacoustical model in frequency domain speech enhancement. IEEE Signal Process. Lett. 11 (2), 270–273.

Jabloun, F., Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 11 (6), 700–708.

Jensen, S.H., Hansen, P.C., Hansen, S.D., Sørensen, J.A., 1995. Reduction of broad-band noise in speech by truncated QSVD. IEEE Trans. Speech Audio Process. 3 (6), 439–448.

Kabal, P., 2002. TSP Speech Database. Technical Report. McGill University. 09(02)

Kırbız, S., Günsel, B., 2013. Perceptually enhanced blind single-channel music source separation by non-negative matrix factorization. Digital Signal Process. 23 (2), 646–658.

Kodrasi, I., Marquardt, D., Doclo, S., 2015. Curvature-based optimization of the trade-off parameter in the speech distortion weighted multichannel Wiener filter. In: IEEE International Conference on Acoustics Speech and Signal Process, Brisbane, Australia, pp. 315–319.

Kwon, K., Shin, J.W., Kim, N.S., 2015. NMF-based speech enhancement using bases update. IEEE Signal Process. Lett. 22 (4), 450–454.

Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. In: Advances in Neural Infomation Processing Systems, pp. 556–562.

Lefevre, A., Bach, F., Févotte, C., 2011. Online algorithms for nonnegative matrix factorization with the itakura-saito divergence. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, USA, pp. 313–316.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Loizou, P.C., 2005. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Process. 13 (5), 857–869.

Mohammadiha, N., Gerkmann, T., Leijon, A., 2011. A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New York, USA, pp. 45–48.

Mohammadiha, N., Smaragdis, P., Leijon, A., 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. IEEE Trans. Audio Speech Lang Process. 21 (10), 2140–2151.

Mysore, G.J., Smaragdis, P., 2011. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. In: IEEE International Conference on Acoustics Speech and Signal Process, Prague, Czech, pp. 17–20.

Natarajan, A., Hansen, J.H., Arehart, K.H., Rossi-Katz, J., 2005. An auditory-masking-threshold-based noise suppression algorithm GMMSE-AMT [ERB] for listeners with sensorineural hearing loss. EURASIP J. Appl. Signal Process. 2005, 2938–2953.

Nikunen, J., Virtanen, T., 2010. Noise-to-mask ratio minimization by weighted non-negative matrix factorization. In: IEEE International Conference on Acoustics Speech and Signal Process, Texas, USA, pp. 25–28.

O'Shaughnessy, D., 1987. Speech Communication: Human and Machine. IEEE Press.

Painter, T., Spanias, A., 2000. Perceptual coding of digital audio. Proc. IEEE 88 (4), 451–515.

Plourde, E., Champagne, B., 2008. Auditory-based spectral amplitude estimators for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1614–1623.

Rangachari, S., Loizou, P.C., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Commun. 48 (2), 220–231.

Recommendation, I., 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation, p. 862.

Scalart, P., Filho, J.V., 1996. Speech enhancement based on a priori signal to noise estimation. In: IEEE International Conference on Acoustics Speech and Signal Process, Atlanta, USA, Vol. 2, pp. 629–632.

Spriet, A., Moonen, M., Wouters, J., 2005. Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel Wiener filtering for noise reduction in hearing aids. IEEE Trans. Signal Process. 53 (3), 911–925.

Varga, A., Steeneken, H.J., 1993. 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 Commun. 12 (3), 247–251.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (2), 126–137.

Virtanen, T., 2007a. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1066–1074.

Virtanen, T., 2007b. Monaural Sound Source Separation by Perceptually Weighted Non-negative Matrix Factorization. Technical Report. Tampere University of Technology..

Virtanen, T., Cemgil, A.T., 2009. Mixtures of gamma priors for non-negative matrix factorization based speech separation. In: Independent Component Analysis and Signal Separation. Springer, pp. 646–653.

You, C.H., Koh, S.N., Rahardja, S., 2005. $\beta$-order MMSE spectral amplitude estimation for speech enhancement. IEEE Trans. Speech Audio Process. 13 (4), 475–486.

Zafeiriou, S., Tefas, A., Buciu, I., Pitas, I., 2006. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. IEEE Trans. Neural Netw. 17 (3), 683–695.