A PRIORI SNR ESTIMATION USING DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION

Ziyi Xu, Samy Elshamy, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig, Germany
{ziyi.xu, s.elshamy, t.fingscheidt}@tu-bs.de

ABSTRACT

A priori signal-to-noise ratio (SNR) contains critical information about the single-channel mixture of a speech and noise signal, and can be used by speech enhancement algorithms. In this paper, we propose a novel *a priori* SNR estimator using the estimates obtained from discriminative non-negative matrix factorization (DNMF). The idea of our new approach is to utilize the DNMF to perform the preliminary speech components estimation, which can be either directly used to estimate the *a priori* SNR, or can be combined with the well-known decision-directed (DD) approach by Ephraim and Malah to perform the *a priori* SNR estimation. We present a speaker-independent but noise-dependent DNMF-based *a priori* SNR estimator. Speech enhancement simulation results in the presence of non-stationary noise validate our new approach combined with well-known spectral weighting rules, outperforming several NMF-based and non-NMF-based state-of-the-art methods, w.r.t. both SNR improvement and speech perceptual quality.

Index Terms— *A priori* SNR, discriminative non-negative matrix factorization, speech enhancement

1. INTRODUCTION

A priori SNR estimation is an important topic in speech enhancement. The estimation task can be very challenging when only a single-channel mixture signal is available. *A priori* SNR estimation methods for single-channel mixtures have been investigated in several publications [1, 2, 3, 4, 5]. A reliable and precise *a priori* SNR estimation is a crucial part in obtaining the spectral weighting rules for speech enhancement algorithms [2, 6, 7]. The decision-directed (DD) method by Ephraim and Malah [1] is a widespread *a priori* SNR estimation approach performing the weighted summation of two components. The first component is the estimated *a priori* SNR obtained from the previous frame's estimated speech signal power and the noise power. The second component is related to the current frame's *a posteriori* SNR.

Non-negative matrix factorization (NMF) and its variant sparse non-negative matrix factorization (SNMF) are popular algorithms used in single-channel speech source separation, such as speech enhancement in the scenario of a speech signal being mixed with non-stationary noise [8, 9]. The basic idea of NMF and SNMF is to represent the non-negative features of every speech source as a product of a codebook matrix for each source and its corresponding weighting matrix. Nevertheless, there are always overlaps of the codebooks between different speech sources which cannot be avoided in general. Then, the codebook's matrices for each source are not discriminative enough to distinguish the elements in the mixed signal. This will degrade the performance of speech enhancement when using conventional NMF or even SNMF.

Discriminative non-negative matrix factorization (DNMF) is a modified SNMF algorithm that is designed to reduce the overlaps between the different sources' codebooks turning out to have better performance than NMF [10, 11]. In the DNMF algorithm proposed in [11], the codebooks for different sources are jointly trained instead of the separate training in NMF.

An algorithm that is using conventional NMF combined with the DD approach to perform the *a priori* SNR estimation and followed by a linear MMSE filter for speech enhancement has been proposed in [12], but still exhibiting and suffering from the issue of too similar codebooks.

In this paper, we propose a new approach that uses DNMF to perform *a priori* SNR estimation, since DNMF performs better than the conventional NMF. We obtain the *a priori* SNR *directly* from the DNMF outputs, or *combine* the DNMF outputs with the DD approach. The estimated *a priori* SNR is evaluated with subsequent well-known spectral weighting rules to perform the actual speech enhancement, such as the well-known Wiener filter (WF) [6], the MMSE log-spectral amplitude estimator (LSA) [2], and the super-Gaussian joint maximum *a posteriori* estimator (SG) [7].

The paper is organized as follows: Section 2 gives a short review of the NMF, SNMF, and DNMF approaches. Section 3 presents our DNMF-based *a priori* SNR estimation methods and the spectral weighting rules to perform the speech enhancement. Simulation and evaluation results are given in Section 4. We conclude this paper in Section 5.

2. BASICS OF DISCRIMINATIVE NON-NEGATIVE MATRIX FACTORIZATION

We assume $y(n)$ to be the observed single-channel mixture of the clean speech signal $s(n)$, and the superimposed noise signal $d(n)$ with n being the discrete-time sample index. In this paper, NMF is used in the discrete Fourier transform (DFT) domain. Therefore, let $Y(\ell,k)$, $S(\ell,k)$, and $D(\ell,k)$ be the respective DFTs, with frame index $\ell \in \mathcal{L} = \{1,2,\ldots,L\}$ and frequency bin index $k \in \mathcal{K} = \{0,1,\ldots,K-1\}$ with DFT size K. Due to the linearity of the DFT, we have

$$Y(\ell,k) = S(\ell,k) + D(\ell,k). \tag{1}$$

As most NMF approaches, we assume that the mixture, the clean speech, and the added noise signal's DFT coefficients have the same phase angles for each frame, so that

$$|Y(\ell,k)| = |S(\ell,k)| + |D(\ell,k)|. \tag{2}$$

Then, we can rewrite the magnitude spectrograms into matrices with dimension $K \times L$ as:

$$\mathbf{Y} = \mathbf{S} + \mathbf{D}. \tag{3}$$

The unknown magnitude spectrograms $\mathbf{S}$ and $\mathbf{D}$ are estimated from $\mathbf{Y}$ by using any of the subsequent NMF variants.

2.1. NMF and SNMF

Non-negative matrix factorization is a method to approximate any non-negative matrix $\mathbf{X}$ by the product of two non-negative matrices, the $K \times N_c$ codebook matrix $\mathbf{B}$ and the $N_c \times L$ weighting matrix $\mathbf{G}$ according to

$$\mathbf{X} \approx \mathbf{BG}, \tag{4}$$

with N_c being the codebook size. The codebook $\mathbf{B}$ and the weighting matrix $\mathbf{G}$ can be obtained by minimizing a cost function as:

$$\min_{\mathbf{B},\mathbf{G}} C\left(\mathbf{X}, \mathbf{BG}\right). \tag{5}$$

In this paper, we investigate different cost functions such as the well-known Kullback-Leibler (KL) divergence C_{KL} [9] and the least squares (LS) C_{LS} [13]. The solution for (5) using the KL divergence can be obtained by multiplicative update rules introduced in [8], while the LS solution can be taken from [13].

In conventional NMF, one often gets a sparse solution for the weighting matrix $\mathbf{G}$, which means that only a subset of the N_c column vectors in $\mathbf{B}$ is used to represent the non-negative matrix $\mathbf{X}$ [10]. To further nourish the sparseness of the solution, sparse non-negative matrix factorization (SNMF) has been proposed [14]. SNMF enforces the sparsity by adding an L_1 norm of the weighting matrix (called $\widetilde{\mathbf{G}}$ here), weighted by μ as a sparseness constraint to (5), and μ is selected from [14]. For SNMF, only the KL divergence is considered in this paper:

$$\min_{\widetilde{\mathbf{B}},\widetilde{\mathbf{G}}} \left(C_{\mathrm{KL}}\left(\mathbf{X}, \widetilde{\mathbf{B}}\widetilde{\mathbf{G}}\right) + \mu \left\|\widetilde{\mathbf{G}}\right\|_1\right). \tag{6}$$

$\widetilde{\mathbf{B}}$ is the codebook matrix with normalized column vectors $\tilde{\mathbf{b}}_j = \left[\frac{b_{1,j}}{\|\mathbf{b}_j\|_1} \quad \cdots \quad \frac{b_{K,j}}{\|\mathbf{b}_j\|_1}\right]^T$, and $\widetilde{\mathbf{G}}$ is the corresponding sparse solution. In the SNMF training stage, we use the multiplicative update rules introduced in [10, 15] to obtain $\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{G}}$ for the speech and noise magnitude spectrograms separately as:

$$\mathbf{S}^{\text{train}} \approx \widetilde{\mathbf{B}}_{\mathrm{s}}^{\text{train}} \widetilde{\mathbf{G}}_{\mathrm{s}}^{\text{train}} \tag{7}$$

and

$$\mathbf{D}^{\text{train}} \approx \widetilde{\mathbf{B}}_{\mathrm{d}}^{\text{train}} \widetilde{\mathbf{G}}_{\mathrm{d}}^{\text{train}}. \tag{8}$$

In the SNMF separation stage, the observed magnitude spectrogram $\mathbf{Y}^{\text{test}}$ is decomposed by fixing the concatenated source codebooks and only using the update rule to obtain the weighting matrix:

$$\begin{aligned} \mathbf{Y}^{\text{test}} &\approx \left[\widetilde{\mathbf{B}}_{\mathrm{s}}^{\text{train}} \quad \widetilde{\mathbf{B}}_{\mathrm{d}}^{\text{train}}\right] \cdot \widetilde{\mathbf{G}}_{\text{mixture}} \\ &= \left[\widetilde{\mathbf{B}}_{\mathrm{s}}^{\text{train}} \quad \widetilde{\mathbf{B}}_{\mathrm{d}}^{\text{train}}\right] \cdot \left[\widetilde{\mathbf{G}}_{\mathrm{s}}^{T} \quad \widetilde{\mathbf{G}}_{\mathrm{d}}^{T}\right]^T, \end{aligned} \tag{9}$$

with $\widetilde{\mathbf{G}}_{\mathrm{s}}$ and $\widetilde{\mathbf{G}}_{\mathrm{d}}$ being the submatrices of $\widetilde{\mathbf{G}}_{\text{mixture}}$. The preliminary estimates for the speech and the noise magnitude spectrograms are calculated by:

$$\breve{\mathbf{S}} = \widetilde{\mathbf{B}}_{\mathrm{s}}^{\text{train}} \widetilde{\mathbf{G}}_{\mathrm{s}} \tag{10}$$

and

$$\breve{\mathbf{D}} = \widetilde{\mathbf{B}}_{\mathrm{d}}^{\text{train}} \widetilde{\mathbf{G}}_{\mathrm{d}}. \tag{11}$$

The final magnitude spectrogram estimates are typically calculated by Wiener filtering:

$$\bar{\mathbf{S}} = \widetilde{\mathbf{H}} \odot \mathbf{Y}^{\text{test}} \tag{12}$$

$$\bar{\mathbf{D}} = \left(\mathbf{I} - \widetilde{\mathbf{H}}\right) \odot \mathbf{Y}^{\text{test}}, \tag{13}$$

with

$$\widetilde{\mathbf{H}} = \left(\breve{\mathbf{S}} \odot \breve{\mathbf{S}}\right) \oslash \left(\breve{\mathbf{S}} \odot \breve{\mathbf{S}} + \breve{\mathbf{D}} \odot \breve{\mathbf{D}}\right) \tag{14}$$

where the operators $\odot$ and $\oslash$, are element-wise product and division, respectively, and $\mathbf{I}$ being the $K \times L^{\text{test}}$ matrix with all entries equal to one.

2.2. Discriminative NMF

In the training stage of SNMF and NMF, the codebook matrices for each source are trained independently as shown in (7) and (8). The cost functions for (7) to (9) ensure the good recovery from the products of the codebooks and the corresponding weighting matrices but do not guarantee the good estimation of each source in the separation stage. In most cases, there are some overlaps in the codebook matrices for each source signal, so that the weighting matrix $\widetilde{\mathbf{G}}_{\text{mixture}}$ obtained in the separation stage does not provide perfect source signal seperation.

A new algorithm called discriminative NMF (DNMF) was proposed to decrease the overlaps of the codebooks between different sources by considering the separation stage in the training stage [11]. The new cost function to obtain the discriminative codebooks $\hat{\mathbf{B}}_{\mathrm{s}}$ and $\hat{\mathbf{B}}_{\mathrm{d}}$ $\left(= \hat{\mathbf{B}}_{\mathrm{s}}^{\text{train}}, \hat{\mathbf{B}}_{\mathrm{d}}^{\text{train}}\right)$ is:

$$\min_{\hat{\mathbf{B}}_{\mathrm{s}},\hat{\mathbf{B}}_{\mathrm{d}}} C\left(\mathbf{S}^{\text{train}}, \widetilde{\mathbf{H}} \odot \mathbf{Y}^{\text{train}}\right), \tag{15}$$

where $\widetilde{\mathbf{H}}$ is then from the SNMF separation stage (9) to (14), but with $\mathbf{Y}^{\text{train}} = \mathbf{S}^{\text{train}} + \mathbf{D}^{\text{train}}$ instead of $\mathbf{Y}^{\text{test}}$ in (9). Thus, the DNMF training is based on the SNMF training without requiring any additional information, which is called parallel training [11]. In (15), the cost function can be either KL divergence or least squares. Then, we can define the two types of DNMF as **DNMF-KL** and **DNMF-LS**, respectively.

In the DNMF training stage, we fix the weighting matrices obtained from parallel SNMF training and only optimize the codebooks by using the multiplicative update rules explained in [11]. The DNMF separation stage is almost the same as the SNMF separation stage when using (9) to (14), but utilizes the discriminative codebooks $\hat{\mathbf{B}}_{\mathrm{s}}^{\text{train}}$ and $\hat{\mathbf{B}}_{\mathrm{d}}^{\text{train}}$ instead of $\widetilde{\mathbf{B}}_{\mathrm{s}}^{\text{train}}$ and $\widetilde{\mathbf{B}}_{\mathrm{d}}^{\text{train}}$ in (10) and (11).

3. NEW DNMF-BASED A PRIORI SNR ESTIMATION

In this paper, we propose a new *a priori* SNR estimation method based on the magnitude spectrograms estimated from DNMF. We define the *a priori* SNR as:

$$\xi\left(\ell,k\right) = \frac{\sigma_S^2\left(\ell,k\right)}{\sigma_D^2\left(\ell,k\right)}, \tag{16}$$

and the *a posteriori* SNR as:

$$\gamma\left(\ell,k\right) = \frac{\left|Y\left(\ell,k\right)\right|^2}{\sigma_D^2\left(\ell,k\right)}, \tag{17}$$

with the entities $\sigma_S^2\left(\ell,k\right)$ and $\sigma_D^2\left(\ell,k\right)$ being the speech and noise signal powers, respectively, which are not available in reality. Instead, we obtain estimations from the magnitude spectrograms of DNMF (12), (13), as

$$\hat{\sigma}_S^2\left(\ell,k\right) = \bar{S}_{i=k,j=\ell}^2 \tag{18}$$

$$\hat{\sigma}_D^2\left(\ell,k\right) = \bar{D}_{i=k,j=\ell}^2. \tag{19}$$

In this work, we do *not* use the DNMF estimates to directly obtain the desired clean speech spectrum. Instead, as in [16], we propose a new *a priori* SNR estimator and then apply a spectral weighting rule $G\left(\ell,k\right)$ to the mixed signal by

$$\hat{S}\left(\ell,k\right) = Y\left(\ell,k\right) \cdot G\left(\ell,k\right). \tag{20}$$

Then, the estimated clean speech spectrum $\hat{S}\left(\ell,k\right)$ is transformed to the time domain by IFFT with overlap add (OLA). For evaluation in Section 4, we will use several well-known methods to calculate spectral weighting rules, such as WF, LSA, and SG, as introduced before. Most spectral weighting rules

$$G\left(\ell,k\right) = f\left(\xi\left(\ell,k\right), \gamma\left(\ell,k\right)\right) \tag{21}$$

are nonlinear functions of the *a priori* SNR and sometimes of the *a posteriori* SNR. In the following, we investigate two different methods to estimate the *a priori* SNR, while the *a posteriori* SNR will be estimated by simply using (19) in the denominator of (17).

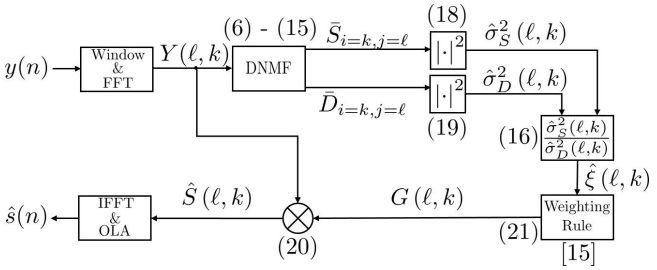

Fig. 1. The new **DNMF-SN** speech enhancement system

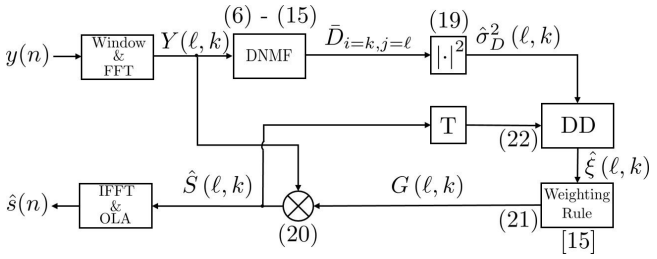

Fig. 2. The new **DNMF-N** speech enhancement system

3.1. DNMF Used for Noise and Speech Power Estimation

As shown in Fig. 1, the speech and the noise power in (16) are directly calculated from the DNMF estimates by (18) and (19). Then, the estimated *a priori* SNR $\hat{\xi}(\ell,k)$ is obtained from (16). We refer to this method as **DNMF-SN**.

The procedures with corresponding formulae are given in Fig. 1. We will investigate both the KL divergence and least squares as a cost function in (15) and refer to the two methods as **DNMF-KL-SN** and **DNMF-LS-SN**, respectively.

3.2. DNMF Used Only for Noise Power Estimation

As shown in Fig. 2, only the noise signal power is estimated through DNMF by (19). Instead of using (16), the *a priori* SNR estimate is then obtained by an instantaneous DD estimator

$$\hat{\xi}(\ell,k) = \beta \frac{\left|\hat{S}(\ell-1,k)\right|^2}{\hat{\sigma}_D^2(\ell-1,k)} + (1-\beta)\cdot\max\{\hat{\gamma}(\ell,k)-1,\xi_{\min}\}, \quad (22)$$

henceforth dubbed **DNMF-N**. The block "T" in Fig. 2 represents a delay of one frame. The lower threshold $\xi_{\min}$ is set to -14 dB for WF and SG, and -15 dB for LSA. We also investigate the two different cost functions for the DNMF training and denote the methods as **DNMF-KL-N** and **DNMF-LS-N**, respectively. The optimal parameters β for the three weighting rules have been reported in and are taken from [17, p. 74].

4. SIMULATION RESULTS

4.1. Setup and Measures

The used speech data to assess our method is taken from the Grid Corpus [18]. To perform the *speaker-independent* speech codebook training as sometimes done in literature [10, 11], we randomly select 16 speakers, which contain 8 male and 8 female and use 32 sentences per speaker for SNMF and DNMF training. For the evaluation, 4 different speakers are chosen, 2 male and 2 female, with 10 sentences each.

Two types of superimposed noise are used: Pedestrian noise (PED) and café noise (CAFE), which are obtained from CHiME-3 data [19]. To obtain the *noise-dependent* noise codebook as usual in NMF and DNMF literature [8, 10, 11, 12], two different noise codebooks are trained separately for two types of noise in the SNMF and the DNMF training stage. The codebook sizes for speech and noise are both set to $N_c = 256$. Each training set consists of distorted speech signals at 7 different SNR levels ranging from -5 dB to 25 dB with a step size of 5 dB. The SNR level is measured according to ITU P.56 [20]. All the speech and noise signals have a sampling rate of 16 kHz and are transformed to the DFT domain using an FFT with $K = 256$, using a periodic Hann window with 128 samples frame shift.

The evaluation set for each noise type consists of the unseen noise and the evaluation speech signals, with SNRs from -5 dB to 20 dB in 5 dB steps. The evaluations are based on the *filtered* clean speech component $\tilde{s}(n)$ and also the *filtered* noise component $\tilde{d}(n)$, which is often called white-box approach [21]. Note that $\tilde{S}(\ell,k) = G(\ell,k)\cdot S(\ell,k)$ and $\tilde{D}(\ell,k) = G(\ell,k)\cdot D(\ell,k)$. We use the following measures [16]:

1) SNR improvement: $\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$, [dB]
with SNR_{in} and SNR_{out} being measured using ITU P.56 [20].

2) Speech *component* quality (PESQ MOS-LQO) is measured using $s(n)$ as reference signal and the *filtered* clean speech component $\tilde{s}(n)$ as test signal according to [22, 23].

3) Segmental speech-to-speech-distortion ratio:

$$\text{SSDR} = \frac{1}{\mathcal{L}_1}\sum_{\ell\in\mathcal{L}_1}\text{SSDR}_{\text{frame}}(\ell) \qquad [\text{dB}]$$

with $\mathcal{L}_1 \subset \mathcal{L}$, denotes the set of speech active frames [24, 16], and

$$\text{SSDR}_{\text{frame}}(\ell) = \max\left\{\min\left\{\text{SSDR}'(\ell), 30\,\text{dB}\right\}, -10\,\text{dB}\right\},$$

and

$$\text{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum_{n\in\mathcal{N}_\ell} s^2(n)}{\sum_{n\in\mathcal{N}_\ell}\left[\tilde{s}(n+\Delta)-s(n)\right]^2}\right],$$

with $\mathcal{N}_\ell$ denoting the sample indices n in frame ℓ, and Δ is used to perform time alignment for the filtered signal $\tilde{s}(n)$.

4) Segmental noise attenuation:

$$\text{NA} = 10\log_{10}\left[\frac{1}{\mathcal{L}}\sum_{\ell\in\mathcal{L}}\text{NA}_{\text{frame}}(\ell)\right], \qquad [\text{dB}]$$

with

$$\text{NA}_{\text{frame}}(\ell) = \frac{\sum_{n\in\mathcal{N}_\ell} d^2(n)}{\sum_{n\in\mathcal{N}_\ell}\tilde{d}^2(n+\Delta)}.$$

4.2. Simulation Results

We report on PED noise and CAFE noise separately. In each type of noise, the measures are averaged over all speakers and also SNR levels, as shown in Tables 1 and 3. Additionally, Tables 2 and 4 show results for an SNR of -5 dB separately. The baseline methods **LSA**, **SG**, and **WF** are weighting rules using DD and minimum statistics (MS) [25] for *a priori* SNR and noise power estimation, respectively. **DNMF-KL** and **DNMF-LS** are pure DNMF baseline methods using different cost functions KL and LS. In each column, the **three best results** are printed in boldface, while the three worst results are marked with a gray underlying color.

From the results in Tables 1 and 3 it becomes clear, that the proposed **DNMF-LS-N** approach shows a more balanced behavior compared to all five baseline approaches with respect to the four measures. Every baseline obtains at least one (in most cases even two) of the three worst scores among the four measures, while **DNMF-LS-N** in those cases, independent of the noise type, performs always better. Furthermore, **DNMF-LS-N** is among the

Method		Δ SNR	PESQ MOS	SSDR	NA
LSA		3.08	3.40	15.05	7.98
SG		2.73	3.37	**15.33**	8.77
WF		3.85	3.45	13.52	10.31
DNMF-KL		5.08	3.05	9.07	**13.17**
DNMF-LS		**5.50**	3.14	9.47	**13.33**
DNMF-KL-SN	LSA	4.69	**3.62**	13.45	11.02
	SG	4.73	3.46	14.20	11.26
	WF	4.83	3.52	13.25	11.39
DNMF-LS-SN	LSA	5.18	**3.70**	13.71	10.99
	SG	5.20	3.52	14.36	11.24
	WF	5.35	3.60	13.51	11.38
DNMF-KL-N	LSA	4.57	**3.64**	14.80	11.25
	SG	4.85	3.42	**15.20**	11.76
	WF	4.91	3.37	13.60	12.13
DNMF-LS-N	LSA	5.28	**3.70**	14.81	11.55
	SG	**5.53**	3.48	**15.15**	12.06
	WF	**5.83**	3.48	13.54	**12.48**

Table 1. Performance for pedestrian (PED) noise averaged over all SNRs; except PESQ MOS all measures in [dB].

Method		Δ SNR	PESQ MOS	SSDR	NA
LSA		2.22	2.54	**5.86**	8.37
SG		2.02	2.72	5.47	9.48
WF		2.41	2.87	4.42	11.19
DNMF-KL		4.19	2.01	2.92	**15.00**
DNMF-LS		4.56	2.16	3.04	**15.13**
DNMF-KL-SN	LSA	3.97	2.82	5.22	12.48
	SG	4.20	2.54	5.32	12.76
	WF	4.13	2.67	4.96	12.87
DNMF-LS-SN	LSA	4.72	**3.00**	5.24	12.41
	SG	5.01	2.70	5.33	12.71
	WF	4.90	2.85	4.97	12.82
DNMF-KL-N	LSA	4.30	2.80	**5.55**	12.98
	SG	4.72	2.53	5.41	13.50
	WF	4.28	2.59	4.48	13.87
DNMF-LS-N	LSA	**5.61**	**3.02**	**5.48**	13.24
	SG	**6.09**	2.74	5.38	13.71
	WF	**6.11**	**2.88**	4.34	**14.11**

Table 2. Performance for pedestrian (PED) noise at SNR = −5 dB; except PESQ MOS all measures in [dB].

top-three in all four tables in all four measures at least in conjunction with one of the weighting rules. This indicates that the typical trade-off between speech quality and noise attenuation could be successfully reduced by the proposed approach. Furthermore, the results are more independent of the chosen weighting rule compared to the three baselines **LSA**, **SG**, and **WF**, which suggests that the *a priori* SNR estimation is more precise. The remainder of the proposed approaches show a similar performance, except for two measures in each table which also indicates a less sensitive trade-off for those new approaches.

Especially, in the low-SNR conditions shown in Tables 2 and 4, a strong improvement of the proposed versus the baseline approaches can be observed in the measures since the difference between the worst scores, obtained by the baseline approaches and the corresponding scores of the proposed approaches is much larger compared to Tables 1 and 3. As the noise types are non-stationary, the advantage might stem from the more instantaneous fashion of estimating the noise power compared to the baselines LSA, SG,

Method		Δ SNR	PESQ MOS	SSDR	NA
LSA		3.48	3.40	14.43	8.43
SG		3.29	3.38	14.65	9.23
WF		4.30	3.48	12.90	10.53
DNMF-KL		4.75	3.12	9.04	**11.90**
DNMF-LS		**5.19**	3.22	9.61	11.43
DNMF-KL-SN	LSA	4.43	**3.73**	13.78	10.36
	SG	4.44	3.58	14.57	10.61
	WF	4.57	3.64	13.58	10.76
DNMF-LS-SN	LSA	4.90	**3.79**	14.28	10.12
	SG	4.87	3.63	14.97	10.39
	WF	5.07	3.70	14.08	10.59
DNMF-KL-N	LSA	4.17	**3.79**	15.51	10.36
	SG	4.44	3.56	**15.93**	11.03
	WF	4.58	3.52	14.27	**11.47**
DNMF-LS-N	LSA	4.87	**3.84**	**15.65**	10.75
	SG	**5.14**	3.62	**16.03**	11.45
	WF	**5.54**	3.60	14.32	**12.00**

Table 3. Performance for cafe (CAFE) noise averaged over all SNRs; except PESQ MOS all measures in [dB].

Method		Δ SNR	PESQ MOS	SSDR	NA
LSA		2.84	2.60	5.55	9.32
SG		3.17	2.75	5.08	10.05
WF		3.14	2.99	4.19	11.50
DNMF-KL		4.28	2.08	2.96	**13.92**
DNMF-LS		4.46	2.26	3.18	13.01
DNMF-KL-SN	LSA	4.35	3.01	6.01	12.00
	SG	4.51	2.77	6.16	12.32
	WF	4.52	2.87	5.73	12.43
DNMF-LS-SN	LSA	4.78	**3.20**	6.14	11.49
	SG	4.96	2.94	6.31	11.83
	WF	5.00	**3.06**	5.85	11.98
DNMF-KL-N	LSA	4.34	3.04	**6.80**	12.21
	SG	4.83	2.73	6.58	12.91
	WF	4.64	2.73	5.50	**13.29**
DNMF-LS-N	LSA	**5.40**	**3.30**	**6.73**	12.38
	SG	**5.90**	2.99	**6.59**	13.07
	WF	**6.28**	3.02	5.31	**13.60**

Table 4. Performance for cafe (CAFE) noise at SNR = −5 dB; except PESQ MOS all measures in [dB].

and WF. The consistent improvement in speech component quality (PESQ MOS) of the four new methods compared to the baselines **DNMF-KL** and **DNMF-LS** could be explained by the fact that using the noise power estimate from the DNMF as input to the DD *a priori* and *a posteriori* SNR estimation and subsequently in a weighting rule is less prone to estimation errors.

5. CONCLUSIONS

In this paper, we have presented two types of DNMF-based *a priori* SNR estimator: **DNMF-SN** and **DNMF-N**. From our evaluations, the **DNMF-LS-N** *a priori* SNR estimation approach offers more balanced behavior compared to all baseline methods and is more independent of the weighting rules by giving a more precise *a priori* SNR estimation. In the low-SNR condition (SNR=−5 dB), the **DNMF-LS-N** approach shows a 1 to 2 dB higher Δ SNR with a similar or better PESQ MOS than all the baseline methods.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] I. Cohen, "Speech Enhancement Using Super-Gaussian Speech Models and Noncausal A Priori SNR Estimation," *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[4] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[5] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 1740–1744.

[6] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[7] 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, no. 7, pp. 1110–1126, 2005.

[8] E. M. Grais and H. Erdogan, "Single Channel Speech Music Separation Using Nonnegative Matrix Factorization and Spectral Masks," in *Proc. of 17th International Conference on Digital Signal Processing*, Corfu, Greece, Jul. 2011, pp. 1–6.

[9] K. Serap and S. Paris, "An Adaptive Time-frequency Resolution Approach for Non-Negative Matrix Factorization Based Single Channel Sound Source Separation," in *Proc. of ICASSP*, Prague, Czech Republic, Jul. 2011, pp. 253–256.

[10] Z. Wang and F. Sha, "Discriminative Non-Negative Matrix Factorization for Single-Channel Speech Separation," in *Proc. of ICASSP*, Florence, Italy, Jul. 2014, pp. 3749–3753.

[11] F. Weninger, J. L. Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its Application to Single-Channel Source Separation," in *Proc. of Interspeech*, Singapore, Singapore, Sep. 2014, pp. 865–869.

[12] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A New Linear MMSE Filter for Single Channel Speech Enhancement Based on Nonnegative Matrix Factorization," in *Proc. of WASPAA*, New Paltz, NY, USA, Nov. 2011, pp. 45–48.

[13] D. D. Lee and H. S. Seung, "Algorithms for Non-Negative Matrix Factorization," in *Proc. of NIPS*, Denver, CO, USA, Jun. 2001, pp. 556–562.

[14] P.O. Hoyer, "Non-Negative Sparse Coding," in *Proc. of Workshop on Neural Networks for Signal Processing*, Martigny, Switzerland, Nov. 2002, pp. 557–565.

[15] T. T. Vu, B. Bigot, and E. S. Chng, "Combining Non-Negative Matrix Factorization and Deep Neural Networks for Speech Enhancement and Automatic Speech Recognition," in *Proc. of ICASSP*, Shanghai, China, May 2016, pp. 499–503.

[16] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[17] H. Yu, *Post-Filter Optimization for Multichannel Automotive Speech Enhancement*, Ph.D. thesis, Institute for Communications Technology, Technische Universität Braunschweig, Germany, 2013.

[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Jun. 2006.

[19] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. of ASRU*, Scottsdale, AZ, USA, Feb. 2015, pp. 504–511.

[20] International Telecommunication Union, *Objective Measurement of Active Speech Level*, Telecommunication Standardization Sector (ITU-T), Rec. P.56, Dec. 2011.

[21] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[22] International Telecommunication Union, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Telecommunication Standardization Sector (ITU-T), Rec. P.862, Feb. 2001.

[23] International Telecommunication Union, *Wideband Hands-Free Communication in Motor Vehicles*, Telecommunication Standardization Sector (ITU-T), Rec. P.1110, Jan. 2015.

[24] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.

[25] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.