Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization

Nasser Mohammadiha*, *Student Member, IEEE,* Paris Smaragdis, *Member, IEEE,* Arne Leijon, *Member, IEEE*

***Abstract*—Reducing the interference noise in a monaural noisy speech signal has been a challenging task for many years. Compared to traditional unsupervised speech enhancement methods, e.g., Wiener filtering, supervised approaches, such as algorithms based on hidden Markov models (HMM), lead to higher-quality enhanced speech signals. However, the main practical difficulty of these approaches is that for each noise type a model is required to be trained a priori. In this paper, we investigate a new class of supervised speech denoising algorithms using nonnegative matrix factorization (NMF). We propose a novel speech enhancement method that is based on a Bayesian formulation of NMF (BNMF). To circumvent the mismatch problem between the training and testing stages, we propose two solutions. First, we use an HMM in combination with BNMF (BNMF-HMM) to derive a minimum mean square error (MMSE) estimator for the speech signal with no information about the underlying noise type. Second, we suggest a scheme to learn the required noise BNMF model online, which is then used to develop an unsupervised speech enhancement system. Extensive experiments are carried out to investigate the performance of the proposed methods under different conditions. Moreover, we compare the performance of the developed algorithms with state-of-the-art speech enhancement schemes using various objective measures. Our simulations show that the proposed BNMF-based methods outperform the competing algorithms substantially.**

***Index Terms*—Nonnegative matrix factorization (NMF), speech enhancement, PLCA, HMM, Bayesian Inference**

I. Introduction

Estimating the clean speech signal in a single-channel recording of a noisy speech signal has been a research topic for a long time and is of interest for various applications including hearing aids, speech/speaker recognition, and speech communication over telephone and internet. A major outcome of these techniques is the improved quality and reduced listening effort in the presence of an interfering noise signal.

In general, speech enhancement methods can be categorized into two broad classes: unsupervised and supervised. Unsupervised methods include a wide range of approaches such as spectral subtraction [1], Wiener and Kalman filtering, e.g., [2], [3], short-time spectral amplitude (STSA) estimators [4], estimators based on super-Gaussian prior distributions for speech DFT coefficients [5]–[8], and schemes based on periodic models of the speech signal [9]. In these methods, a statistical model is assumed for the speech and noise signals, and the clean speech is estimated from the noisy observation without any prior information on the noise type or speaker identity. However, the main difficulty of most of these methods is estimation of the noise power spectral density (PSD) [10]–[12], which is a challenging task if the background noise is non-stationary.

For the supervised methods, a model is considered for both the speech and noise signals and the model parameters are estimated using the training samples of that signal. Then, an interaction model is defined by combining speech and noise models and the noise reduction task is carried out. Some examples of this class of algorithms include the codebook-based approaches, e.g., [13], [14] and hidden Markov model (HMM) based methods [15]–[19]. One advantage of these methods is that there is no need to estimate the noise PSD using a separate algorithm.

The supervised approaches have been shown to produce better quality enhanced speech signals compared to the unsupervised methods [14], [16], which can be expected as more prior information is fed to the system in these cases and the considered models are trained for each specific type of signals. The required prior information on noise type (and speaker identity in some cases) can be given by the user, or can be obtained using a built-in classification scheme [14], [16], or can be provided by a separate acoustic environment classification algorithm [20]. The primary goal of this work is to propose supervised and unsupervised speech enhancement algorithms based on nonnegative matrix factorization (NMF) [21], [22].

NMF is a technique to project a nonnegative matrix $\mathbf{y}$ onto a space spanned by a linear combination of a set of basis vectors, i.e., $\mathbf{y} \approx \mathbf{bv}$, where both $\mathbf{b}$ and $\mathbf{v}$ are nonnegative matrices. In speech processing, $\mathbf{y}$ is usually the spectrogram of the speech signal with spectral vectors stored by column, $\mathbf{b}$ is the basis matrix or dictionary, and $\mathbf{v}$ is referred to as the NMF coefficient or activation matrix. NMF has been widely used as a source separation technique applied to monaural mixtures, e.g., [23]–[25]. More recently, NMF has also been used to estimate the clean speech from a noisy observation [26]–[31].

When applied to speech source separation, a good separation can be expected only when speaker-dependent basis are learned. In contrast, for noise reduction, even if a general speaker-independent basis matrix of speech is learned, a good enhancement can be achieved [29], [31]. Nevertheless, there might be some scenarios (such as speech degraded with

N. Mohammadiha and A. Leijon are with the Department of Electrical Engineering, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden (e-mail: nmoh@kth.se; leijon@kth.se).

P. Smaragdis is with the Department of Computer Science and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, IL, USA (e-mail: paris@illinois.edu).

multitalker babble noise) for which the basis matrices of speech and noise are quite similar. In these cases, although the traditional NMF-based approaches can be used to get state-of-the-art performance, other constraints can be imposed into NMF to obtain a better noise reduction. For instance, assuming that the babble waveform is obtained as a sum of different speech signals, a nonnegative hidden Markov model is proposed in [26] to model the babble noise in which the babble basis is identical to the speech basis. Another fundamental issue in basic NMF is that it ignores the important temporal dependencies of the audio signals. Different approaches have been proposed in the literature to employ temporal dynamics in NMF, e.g., [23]–[25], [27], [30], [31].

In this paper, we first propose a new supervised NMF-based speech enhancement system. In the proposed method, the temporal dependencies of speech and noise signals are used to construct informative prior distributions that are applied in a Bayesian framework to perform NMF (BNMF). We then develop an HMM structure with output density functions given by BNMF to simultaneously classify the environmental noise and enhance the noisy signal. Therefore, the noise type doesn't need to be specified a priori. Here, the classification is done using the noisy input and is not restricted to be applied at only the speech pauses as it is in [16], and it doesn't require any additional noise PSD tracking algorithm, as it is required in [14].

Moreover, we propose an unsupervised NMF-based approach in which the noise basis matrix is learned online from the noisy mixture. Although online dictionary learning from clean data has been addressed in some prior works, e.g., [32], [33], our causal method learns the noise basis matrix from the noisy mixture. The main contributions of this work can be summarized as:

1) We present a review of state-of-the-art NMF-based noise reduction approaches.
2) We propose a speech enhancement method based on BNMF that inherently captures the temporal dependencies in the form of hierarchical prior distributions. Some preliminary results of this approach has been presented in [31]. Here, we further develop the method and evaluate its performance comprehensively. In particular, we present an approach to construct SNR-dependent prior distributions.
3) An environmental noise classification technique is suggested and is combined with the above BNMF approach (BNMF-HMM) to develop an unsupervised speech enhancement system.
4) A causal online dictionary learning scheme is proposed that learns the noise basis matrix from the noisy observation. Our simulations show that the final unsupervised noise reduction system outperforms state-of-the-art approaches significantly.

The rest of the paper is organized as follows: The review of the NMF-based speech enhancement algorithms is presented in Section II. In Section III, we describe our main contributions, namely the BNMF-based noise reduction, BNMF-HMM structure, and online noise dictionary learning. Section IV presents our experiments and results with supervised and unsupervised noise reduction systems. Finally, Section V concludes the study.

TABLE I
THE TABLE SUMMARIZES SOME OF THE NOTATIONS THAT ARE CONSISTENTLY USED IN THE PAPER.

k	frequency index
t	time index
X	a scalar random variable
$\mathbf{Y}=[Y_{kt}]$	a matrix of random variabels
$\mathbf{Y}_t$	t-th column of $\mathbf{Y}$
$\mathbf{y}=[y_{kt}]$	a matrix of observed magnitude spectrogram
$\mathbf{y}_t$	t-th column of $\mathbf{y}$
$\mathbf{b}^{(s)}$	speech parameters ($\mathbf{b}^{(s)}$ is the speech basis matrix)
$\mathbf{b}^{(n)}$	noise parameters ($\mathbf{b}^{(n)}$ is the noise basis matrix)
$\mathbf{b}=\left[\mathbf{b}^{(s)}\ \mathbf{b}^{(n)}\right]$	mixture parameters ($\mathbf{b}$ is the mixture basis matrix)

II. Review of State-of-the-art NMF-Based Speech Enhancement

In this section, we first explain a basic NMF approach, and then we review NMF-based speech enhancement. Let us represent the random variables associated with the magnitude of the discrete Fourier transform (DFT) coefficients of the speech, noise, and noisy signals as $\mathbf{S}=[S_{kt}]$, $\mathbf{N}=[N_{kt}]$ and $\mathbf{Y}=[Y_{kt}]$, respectively, where k and t denote the frequency and time indices, respectively. The actual realizations are shown in small letters, e.g., $\mathbf{y}=[y_{kt}]$. Table I summarizes some of the notations that are frequently used in the paper.

To obtain a nonnegative decomposition of a given matrix $\mathbf{x}$, a cost function is usually defined and is minimized. Let us denote the basis matrix and NMF coefficient matrix by $\mathbf{b}$ and $\mathbf{v}$, respectively. Nonnegative factorization is achieved by solving the following optimization problem:

$$(\mathbf{b},\mathbf{v}) \quad = \quad \underset{\mathbf{b},\mathbf{v}}{\arg\min} \ \ D(\mathbf{y}\|\mathbf{bv})+\mu h\left(\mathbf{b},\mathbf{v}\right), \qquad (1)$$

where $D(\mathbf{y}\|\hat{\mathbf{y}})$ is a cost function, $h(\cdot)$ is an optional regularization term, and μ is the regularization weight. The minimization in (1) is performed under the nonnegativity constraint of $\mathbf{b}$ and $\mathbf{v}$. The common choices for the cost function include Euclidean distance [21], generalized Kullback-Leibler divergence [21], [34], Itakura-Saito divergence [25], and the negative likelihood of data in the probabilistic NMFs [35]. Depending on the application, the sparsity of the activations $\mathbf{v}$ and the temporal dependencies of input data $\mathbf{x}$ are two popular motivations to design the regularization function, e.g., [24], [27], [36], [37]. Since (1) is not a convex problem, iterative gradient descent or expectation-maximization (EM) algorithms are usually followed to obtain a locally optimal solution for the problem [21], [25], [35].

Let us consider a supervised denoising approach where the basis matrix of speech $\mathbf{b}^{(s)}$ and the basis matrix of noise $\mathbf{b}^{(n)}$ are learned using the appropriate training data in advance. The common assumption used to model the noisy speech signal is the additivity of speech and noise spectrograms, i.e., $\mathbf{y}=\mathbf{s}+\mathbf{n}$. Although in the real world problems this assumption is not justified completely, the developed algorithms have been shown to produce satisfactory results, e.g., [24]. The basis matrix of

the noisy signal is obtained by concatenating the speech and noise basis matrices as $\mathbf{b}=[\mathbf{b}^{(s)}\,\mathbf{b}^{(n)}]$. Given the magnitude of DFT coefficients of the noisy speech at time t, $\mathbf{y}_t$, the problem in (1) is now solved—with $\mathbf{b}$ held fixed—to obtain the noisy NMF coefficients $\mathbf{v}_t$. The NMF decomposition takes the form $\mathbf{y}_t \approx \mathbf{b}\mathbf{v}_t = [\mathbf{b}^{(s)}\,\mathbf{b}^{(n)}][(\mathbf{v}_t^{(s)})^\top\ (\mathbf{v}_t^{(n)})^\top]^\top$, where $\top$ denotes transposition. Finally, an estimate of the clean speech DFT magnitudes is obtained by a Wiener-type filtering as:

$$\hat{\mathbf{s}}_t = \frac{\mathbf{b}^{(s)}\mathbf{v}_t^{(s)}}{\mathbf{b}^{(s)}\mathbf{v}_t^{(s)} + \mathbf{b}^{(n)}\mathbf{v}_t^{(n)}} \odot \mathbf{y}_t, \tag{2}$$

where the division is performed element-wise, and $\odot$ denotes an element-wise multiplication. The clean speech waveform is estimated using the noisy phase and inverse DFT. One advantage of the NMF-based approaches over the HMM-based [16], [17] or codebook-driven [14] approaches is that NMF automatically captures the long-term levels of the signals, and no additional gain modeling is necessary.

Schmidt *et al.* [28] presented an NMF-based unsupervised batch algorithm for noise reduction. In this approach, it is assumed that the entire noisy signal is observed, and then the noise basis vectors are learned during the speech pauses. In the intervals of speech activity, the noise basis matrix is kept fixed and the rest of the parameters (including speech basis and speech and noise NMF coefficients) are learned by minimizing the Euclidean distance with an additional regularization term to impose sparsity on the NMF coefficients. The enhanced signal is then obtained similarly to (2). The reported results show that this method outperforms a spectral subtraction algorithm, especially for highly non-stationary noises. However, the NMF approach is sensitive to the performance of the voice activity detector (VAD). Moreover, the proposed algorithm in [28] is applicable only in the batch mode, which is usually not practical in the real world.

In [27], a supervised NMF-based denoising scheme is proposed in which a heuristic regularization term is added to the cost function. By doing so, the factorization is enforced to follow the pre-obtained statistics. In this method, the basis matrices of speech and noise are learned from training data offline. Also, as part of the training, the mean and covariance of the log of the NMF coefficients are computed. Using these statistics, the negative likelihood of a Gaussian distribution (with the calculated mean and covariance) is used to regularize the cost function during the enhancement. The clean speech signal is then estimated as $\hat{\mathbf{s}}_t = \mathbf{b}^{(s)}\mathbf{v}_t^{(s)}$. Although it is not explicitly mentioned in [27], to make regularization meaningful the statistics of the speech and noise NMF coefficients have to be adjusted according to the long-term levels of speech and noise signals.

In [29], authors propose a linear minimum mean square error (MMSE) estimator for NMF-based speech enhancement. In this work, NMF is applied to $\mathbf{y}_t^p$ (i.e., $\mathbf{y}_t^p = \mathbf{b}\mathbf{v}_t$, where $p = 1$ corresponds to using magnitude of DFT coefficients and $p = 2$ corresponds to using magnitude-squared DFT coefficients) in a frame by frame routine. Then, a gain variable $\mathbf{g}_t$ is estimated to filter the noisy signal as: $\hat{\mathbf{s}}_t = (\mathbf{g}_t \odot \mathbf{y}_t^p)^{1/p}$. Assuming that the basis matrices of speech and noise are obtained during the training stage, and that the NMF coefficients $\mathbf{V}_t$ are random variables, $\mathbf{g}_t$ is derived such that the mean square error between $\mathbf{S}_t^p$ and $\widehat{\mathbf{S}_t^p}$ is minimized. The optimal gain is shown to be:

$$\mathbf{g}_t = \frac{\boldsymbol{\xi}_t + c^2\sqrt{\boldsymbol{\xi}_t}}{\boldsymbol{\xi}_t + 1 + 2c^2\sqrt{\boldsymbol{\xi}_t}}, \tag{3}$$

where c is a constant that depends on p [29] and $\boldsymbol{\xi}_t$ is called the smoothed speech to noise ratio that is estimated using a decision-directed approach. For a theoretical comparison of (3) to a usual Wiener filter see [29]. The conducted simulations show that the results using $p = 1$ are superior to those using $p = 2$ (which is in line with previously reported observations, e.g., [24]) and that both of them are better than the results of a state-of-the-art Wiener filter.

A semi-supervised approach is proposed in [30] to denoise a noisy signal using NMF. In this method, a nonnegative hidden Markov model (NHMM) is used to model the speech magnitude spectrogram. Here, the HMM state-dependent output density functions are assumed to be a mixture of multinomial distributions, and thus, the model is closely related to probabilistic latent component analysis (PLCA) [35]. An NHMM is described by a set of basis matrices and a Markovian transition matrix that captures the temporal dynamics of the underlying data. To describe a mixture signal, the corresponding NHMMs are then used to construct a factorial HMM. When applied for noise reduction, first a speaker-dependent NHMM is trained on a speech signal. Then, assuming that the whole noisy signal is available (batch mode), the EM algorithm is run to simultaneously estimate a single-state NHMM for noise and also to estimate the NMF coefficients of the speech and noise signals. The proposed algorithm doesn't use a VAD to update the noise dictionary, as was done in [28]. But the algorithm requires the entire spectrogram of the noisy signal, which makes it difficult for practical applications. Moreover, the employed speech model is speaker-dependent, and requires a separate speaker identification algorithm in practice. Finally, similar to the other approaches based on the factorial models, the method in [30] suffers from high computational complexity.

A linear nonnegative dynamical system is presented in [38] to model temporal dependencies in NMF. The proposed causal filtering and fixed-lag smoothing algorithms use Kalman-like prediction in NMF and PLCA. Compared to the ad-hoc methods that use temporal correlations to design regularity functions, e.g., [27], [37], this approach suggests a solid framework to incorporate temporal dynamics into the system. Also, the computational complexity of this method is significantly less than [30].

Raj *et al.* [39] proposed a phoneme-dependent approach to use NMF for speech enhancement in which a set of basis vectors are learned for each phoneme a priori. Given the noisy recording, an iterative NMF-based speech enhancer combined with an automatic speech recognizer (ASR) is pursued to estimate the clean speech signal. In the experiments, a mixture of speech and music is considered and using a set of speaker-dependent basis matrices the estimation of the clean speech is carried out.

NMF-based noise PSD estimation is addressed in [37]. In this work, the speech and noise basis matrices are trained

offline, after which a constrained NMF is applied to the noisy spectrogram in a frame by frame basis. To utilize the time dependencies of the speech and noise signals, an l_2-norm regularization term is added to the cost function. This penalty term encourages consecutive speech and noise NMF coefficients to take similar values, and hence, to model the signals' time dependencies. The instantaneous noise periodogram is obtained similarly to (2) by switching the role of speech and noise approximates. This estimate is then smoothed over time using an exponential smoothing to get a less-fluctuating estimate of the noise PSD, which can be combined with any algorithm that needs a noise PSD, e.g., Wiener filter.

III. Speech Enhancement Using Bayesian NMF

In this section, we present our Bayesian NMF (BNMF) based speech enhancement methods. In the following, an overview of the employed BNMF is provided first, which was originally proposed in [34]. Our proposed extensions of this BNMF to modeling a noisy signal, namely BNMF-HMM and Online-BNMF are given in Subsections III-A and III-B, respectively. Subsection III-C presents a method to construct informative priors to use temporal dynamics in NMF.

The probabilistic NMF in [34] assumes that an input matrix is stochastic, and to perform NMF as $\mathbf{y} \approx \mathbf{bv}$ the following model is considered:

$$Y_{kt} = \sum_i Z_{kit}, \tag{4}$$

$$\begin{aligned} f_{Z_{kit}}\left(z_{kit}\right) &= \mathcal{PO}\left(z_{kit}; b_{ki}v_{it}\right) \\ &= \left(b_{ki}v_{it}\right)^{z_{kit}} e^{-b_{ki}v_{it}} / \left(z_{kit}!\right), \end{aligned} \tag{5}$$

where Z_{kit} are latent variables, $\mathcal{PO}\left(z;\lambda\right)$ denotes the Poisson distribution, and $z!$ is the factorial of z. A schematic representation of this model is shown in Fig. 1.

As a result of (4) and (5), Y_{kt} is assumed Poisson-distributed and integer-valued. In practice, the observed spectrogram is first scaled up and then rounded to the closest integer numbers to avoid large quantization errors. The maximum likelihood (ML) estimate of the parameters $\mathbf{b}$ and $\mathbf{v}$ can be obtained using an EM algorithm [34], and the result would be identical to the well-known multiplicative update rules for NMF using Kullback-Leibler (KL-NMF) divergence [21].

In the Bayesian formulation, the nonnegative factors are further assumed to be random variables. In this hierarchical model, gamma prior distributions are considered to govern the basis ($\mathbf{B}$) and NMF coefficient ($\mathbf{V}$) matrices:

$$\begin{aligned} f_{V_{it}}\left(v_{it}\right) &= \mathcal{G}\left(v_{it}; \phi_{it}, \theta_{it}/\phi_{it}\right), \\ f_{B_{ki}}\left(b_{ki}\right) &= \mathcal{G}\left(b_{ki}; \psi_{ki}, \gamma_{ki}/\psi_{ki}\right), \end{aligned} \tag{6}$$

in which $\mathcal{G}\left(v;\phi,\theta\right) = \exp((\phi-1)\log v - v/\theta - \log\Gamma\left(\phi\right) - \phi\log\theta)$ denotes the gamma density function with ϕ as the shape parameter and θ as the scale parameter, and $\Gamma\left(\phi\right)$ is the gamma function. ϕ, θ, ψ, and γ are referred to as the hyperparameters.

As the exact Bayesian inference for (4), (5), and (6) is difficult, a variational Bayes approach has been proposed in [34] to obtain the approximate posterior distributions of $\mathbf{B}$ and $\mathbf{V}$. In this approximate inference, it is assumed that the posterior distribution of the parameters are independent, and these uncoupled posteriors are inferred iteratively by maximizing a lower bound on the marginal log-likelihood of data.

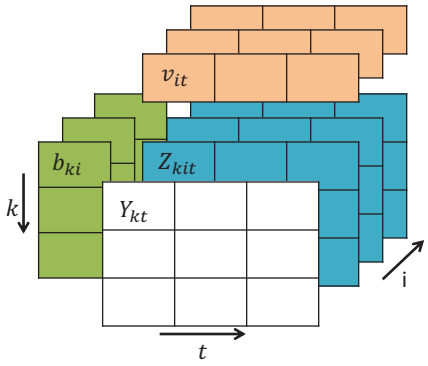

Fig. 1. A schematic representation of (4) and (5) [34]. Each time-frequency bin of a magnitude spectrogram (Y_{kt}) is assumed to be a sum of some Poisson-distributed hidden random variables (Z_{kit}).

More specifically for this Bayesian NMF, in an iterative scheme, the current estimates of the posterior distributions of $\mathbf{Z}$ are used to update the posterior distributions of $\mathbf{B}$ and $\mathbf{V}$, and these new posteriors are used to update the posteriors of $\mathbf{Z}$ in the next iteration. The iterations are carried on until convergence. The posterior distributions for $Z_{k,:,t}$ are shown to be multinomial density functions (: denotes 'all the indices'), while for B_{ki} and V_{it} they are gamma density functions. Full details of the update rules can be found in [34]. This variational approach is much faster than an alternative Gibbs sampler, and its computational complexity can be comparable to that of the ML estimate of the parameters (KL-NMF).

A. BNMF-HMM for Simultaneous Noise Classification and Reduction

In the following, we describe the proposed BNMF-HMM noise reduction scheme in which the state-dependent output density functions are instances of the BNMF explained in the introductory part of this section. Each state of the HMM corresponds to one specific noise type. Let us consider a set of noise types for which we are able to gather some training data, and let us denote the cardinality of the set by M. We can train a BNMF model for each of these noise types given its training data. Moreover, we consider a universal BNMF model for speech that can be trained a priori. Note that the considered speech model doesn't introduce any limitation in the method since we train a model for the speech signal in general, and we don't use any assumption on the identity or gender of the speakers.

The structure of the BNMF-HMM is shown in Fig. 2. Each state of the HMM has some state-dependent parameters, which are the noise BNMF model parameters. Also, all the states share some state-independent parameters, which consist of the speech BNMF model and an estimate of the long-term signal to noise ratio (SNR) that will be used for the enhancement. To complete the Markovian model, we need to predefine an empirical state transition matrix (whose dimension is $M \times$

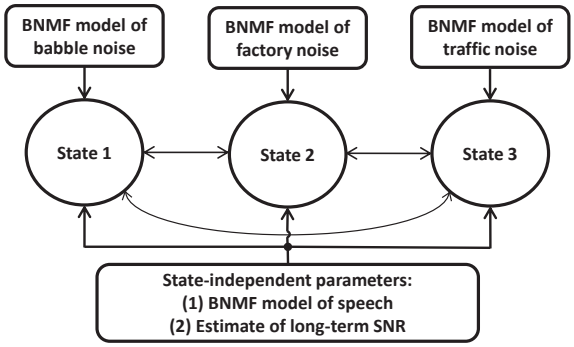

Fig. 2. A block diagram representation of BNMF-HMM with three states.

M) and an initial state probability vector. For this purpose, we assign some high values to the diagonal elements of the transition matrix, and we set the rest of its elements to some small values such that each row of the transition matrix sums to one. Each element of the initial state probability vector is also set to $1/M$.

We model the magnitude spectrogram of the clean speech and noise signals by (4). To obtain a BNMF model, we need to find the posterior distribution of the basis matrix, and optimize for the hyperparameters if desired. During training, we assign some sparse and broad prior distributions to $\mathbf{B}$ and $\mathbf{V}$ according to (6). For this purpose, ψ and γ are chosen such that the mean of the prior distribution for $\mathbf{B}$ is small, and its variance is very high. On the other hand, ϕ and θ are chosen such that the prior distribution of $\mathbf{V}$ has a mean corresponding to the scale of the data and has a high variance to represent uncertainty. To have good initializations for the posterior means, the multiplicative update rules for KL-NMF are applied first for a few iterations, and the result is used as the initial values for the posterior means. After the initialization, variational Bayes (as explained before) is run until convergence. We also optimize the hyperparameters using Newton's method, as proposed in [34].

In the following, the speech and noise random basis matrices are denoted by $\mathbf{B}^{(s)}$ and $\mathbf{B}^{(n)}$, respectively. A similar notation is used to distinguish all the speech and noise parameters.

Let us denote the hidden state variable at each time frame t by X_t, which can take one of the M possible outcomes $x_t = 1, 2, \ldots M$. The noisy magnitude spectrogram, given the state X_t, is modeled using (4). Here, we use the additivity assumption to approximate the state-dependent distribution of the noisy signal, i.e., $\mathbf{y}_t = \mathbf{s}_t + \mathbf{n}_t$. To obtain the distribution of the noisy signal, given the state X_t, the parameters of the speech and noise basis matrices ($\mathbf{B}^{(s)}$ and $\mathbf{B}^{(n)}$) are concatenated to obtain the parameters of the noisy basis matrix $\mathbf{B}$. Since the sum of independent Poisson random variables is Poisson, (4) leads to:

$$f_{Y_{kt}}\left(y_{kt} \mid x_t, \mathbf{b}, \mathbf{v}_t\right) = \frac{\lambda_{kt}^{y_{kt}} e^{-\lambda_{kt}}}{y_{kt}!}\ , \tag{7}$$

where $\lambda_{kt} = \sum_i b_{ki} v_{it}$. Note that although the basis matrix $\mathbf{b}$ is state-dependent, to keep the notations uncluttered, we skip writing this dependency explicitly.

The state-conditional likelihood of the noisy signal can now be computed by integrating over $\mathbf{B}$ and $\mathbf{V}_t$ as:

$$\begin{aligned} f_{Y_{kt}}\left(y_{kt} \mid x_t\right) &= \int\int f_{Y_{kt},\mathbf{B},\mathbf{V}_t}\left(y_{kt}, \mathbf{b}, \mathbf{v}_t \mid x_t\right) d\mathbf{b} d\mathbf{v}_t \\ &= \int\int f_{Y_{kt}}\left(y_{kt} \mid \mathbf{b}, \mathbf{v}_t, x_t\right) \\ &\quad f_{\mathbf{B},\mathbf{V}_t}\left(\mathbf{b}, \mathbf{v}_t \mid x_t\right) d\mathbf{b} d\mathbf{v}_t. \end{aligned} \tag{8}$$

The distribution of $\mathbf{y}_t$ is obtained by assuming that different frequency bins are independent [5], [7]:

$$f_{\mathbf{Y}_t}\left(\mathbf{y}_t \mid x_t\right) = \prod_k f_{Y_{kt}}\left(y_{kt} \mid x_t\right). \tag{9}$$

As the first step of the enhancement, variational Bayes approach is applied to approximate the posterior distributions of the NMF coefficient vector $\mathbf{V}_t$ by maximizing the variational lower bound on (9). Here, we assume that the state-dependent posterior distributions of $\mathbf{B}$ are time-invariant and are identical to those obtained during the training. Moreover, we use the temporal dynamics of noise and speech to construct informative prior distributions for $\mathbf{V}_t$, which is explained in Subsection III-C. After convergence of the variational learning, we will have the parameters (including expected values) of the posterior distributions of $\mathbf{V}_t$ as well as the latent variables $\mathbf{Z}_t$.

The MMSE estimate [40] of the speech DFT magnitudes can be shown to be [15], [26]:

$$\hat{s}_{kt} = E\left(S_{kt} \mid \mathbf{y}_t\right) = \frac{\sum_{x_t=1}^{M} \xi_t\left(\mathbf{y}_t, x_t\right) E\left(S_{kt} \mid x_t, \mathbf{y}_t\right)}{\sum_{x_t=1}^{M} \xi_t\left(\mathbf{y}_t, x_t\right)}, \tag{10}$$

where

$$\begin{aligned} \xi_t\left(\mathbf{y}_t, x_t\right) &= f_{\mathbf{Y}_t, X_t}\left(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1}\right) \\ &= f_{\mathbf{Y}_t}\left(\mathbf{y}_t \mid x_t\right) f_{X_t}\left(x_t \mid \mathbf{y}_1^{t-1}\right), \end{aligned} \tag{11}$$

in which $\mathbf{y}_1^{t-1} = \{\mathbf{y}_1, \ldots \mathbf{y}_{t-1}\}$. Here, $f_{X_t}(x_t \mid \mathbf{y}_1^{t-1})$ is computed using the forward algorithm [41]. Since (8) can not be evaluated analytically, one can either use numerical methods or use approximations to calculate $f_{Y_{kt}}\left(y_{kt} \mid x_t\right)$. Instead of expensive stochastic integrations, we approximate (8) by evaluating the integral at the mean value of the posterior distributions of $\mathbf{B}$ and $\mathbf{V}_t$:

$$f_{Y_{kt}}\left(y_{kt} \mid x_t\right) \approx f_{Y_{kt}}\left(y_{kt} \mid \mathbf{b}', \mathbf{v}_t', x_t\right), \tag{12}$$

where $\mathbf{b}' = E(\mathbf{B} \mid \mathbf{y}_t, x_t)$, and $\mathbf{v}_t' = E(\mathbf{V}_t \mid \mathbf{y}_t, x_t)$ are the posterior means of the basis matrix and NMF coefficient vector that are obtained using variational Bayes. Other types of point approximations have also been used for gain modeling in the context of HMM-based speech enhancement [17], [18].

To finish our derivation, we need to calculate the state-dependent MMSE estimate of the speech DFT magnitudes $E(S_{kt} \mid x_t, \mathbf{y}_t)$. First, let us rewrite (4) for the noisy signal as:

$$Y_{kt} = S_{kt} + N_{kt} = \sum_{i=1}^{I^{(s)}} Z_{kit}^{(s)} + \sum_{i=1}^{I^{(n)}} Z_{kit}^{(n)} = \sum_{i=1}^{I^{(s)}+I^{(n)}} Z_{kit},$$

where $I^{(s)}$ and $I^{(n)}$ are the number of speech and noise basis vectors, respectively, given X_t. Then,

$$E\left(S_{kt} \mid x_t, \mathbf{y}_t\right) = E\left(\sum_{i=1}^{I^{(s)}} Z_{kit}^{(s)} \mid x_t, \mathbf{y}_t\right) = \sum_{i=1}^{I^{(s)}} E\left(Z_{kit}^{(s)} \mid x_t, \mathbf{y}_t\right). \quad (13)$$

The posterior expected values of the latent variables in (13) are obtained during variational Bayes and are given by [34]:

$$E\left(Z_{kit} \mid x_t, \mathbf{y}_t\right) = \frac{e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}}{\sum_{i=1}^{I^{(s)}+I^{(n)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}} y_{kt}. \quad (14)$$

Finally, using (14) in (13), we get

$$E\left(S_{kt} \mid x_t, \mathbf{y}_t\right) = \frac{\sum_{i=1}^{I^{(s)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}}{\sum_{i=1}^{I^{(s)}+I^{(n)}} e^{E(\log B_{ki} + \log V_{it} | x_t, \mathbf{y}_t)}} y_{kt}. \quad (15)$$

As mentioned before, the posterior distributions of $\mathbf{B}$ and $\mathbf{V}$ are gamma density functions and the required expected values to evaluate (15) are available in closed form. The time-domain enhanced speech signal is reconstructed using (10) and the noisy phase information.

Eq. (15) includes Wiener filtering (2) as a special case. When the posterior distributions of the basis and NMF coefficients are very sharp (which happens for large shape parameters in the gamma distribution), $E(\log V_{it} \mid x_t, \mathbf{y}_t)$ approaches the logarithm of the mean value of the posterior distribution, $\log(E(V_{it} \mid x_t, \mathbf{y}_t))$. This can be easily verified by considering that for very large arguments the logarithm provides an accurate approximation to the digamma function. Therefore, for large posterior shape parameters (15) converges asymptotically to (2). In this case, the mean values of the posterior distributions are used to design the Wiener filter.

We can use $\xi_t\left(\mathbf{y}_t, x_t\right)$ to classify the acoustic noise more explicitly. For this purpose, we compute the posterior state probability as:

$$f\left(x_t \mid \mathbf{y}_1^t\right) = \frac{f(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1})}{\sum_{x_t} f(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1})}. \quad (16)$$

To reduce fluctuations, it is helpful to smooth (16) over time. Other likelihood-based classification techniques have been used in [14], [16] for HMM-based and codebook-driven denoising approaches. In [14], a long-term noise PSD is computed using a separate noise PSD tracking algorithm and is used to select one of the available noise models to enhance the noisy signal. Alternatively, in [16], a single noise HMM is selected during periods of speech pauses and is used to enhance the noisy signal until the next speech pause when a new selection is made. Our proposed classification in (16) neither needs an additional noise PSD tracking algorithm, nor requires a voice activity detector.

B. Online Noise Basis Learning for BNMF

We present our scheme to learn the noise basis matrix from the noisy data in this subsection. The online-adapted noise basis is then employed to enhance the noisy signal using the BNMF approach, similarly to III-A with only one state in the HMM. We continue to use a universal speech model that is learned offline.

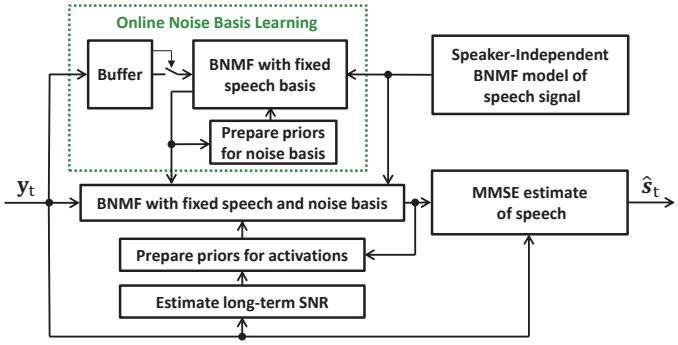

Fig. 3. Block diagram representation of BNMF with online noise basis learning. $\mathbf{y}_t$ and $\hat{\mathbf{s}}_t$ are the short-time spectral amplitudes of the noisy and enhanced speech signals, receptively, at time frame t. The goal of the "Prepare priors" boxes is to recursively update the prior distributions, which will be also discussed in III-C.

To update the noise basis, we store N_1 past noisy magnitude DFT frames into a buffer $\underline{\mathbf{n}} \in \mathbb{R}_+^{K \times N_1}$, where K is the length of $\mathbf{y}_t$. The buffer will be updated when a new noisy frame arrives. Then, keeping the speech basis unchanged, variational Bayes is applied to $\underline{\mathbf{n}}$ to find the posterior distributions of both the speech and noise NMF coefficients and noise basis matrix.

Let us denote the noise dictionary at time index $t-1$ by $f_{\mathbf{B}_{t-1}^{(n)}}(\mathbf{b}_{t-1}^{(n)} \mid \mathbf{y}_1^{t-1})$. To maintain a slowly varying basis matrix, we flatten $f_{\mathbf{B}_{t-1}^{(n)}}(\mathbf{b}_{t-1}^{(n)} \mid \mathbf{y}_1^{t-1})$ and use it as the prior distribution for the noise basis matrix at time t. Accordingly, using the notation from (6), we set $\boldsymbol{\gamma}^{(n)} = E(\mathbf{B}_t^{(n)}) = E\left(\mathbf{B}_{t-1}^{(n)} \mid \mathbf{y}_1^{t-1}\right)$, and $\psi_{ki}^{(n)}$ is set to a high value ($\psi_{ki}^{(n)} = \psi^{(n)} \gg 1, k = 1, \ldots K, i = 1, \ldots I^{(n)}$) to avoid overfitting. With a high value for the shape parameter, the posterior distributions are flattened only slightly to obtain a quite sharp prior distribution. Therefore, the posteriors of the noise basis matrix are encouraged to follow the prior patterns unless the noise spectrogram changes heavily. Fig. 3 shows a simplified diagram of the online BNMF approach. The top part of the figure (dashed-line box) illustrates the online noise basis learning.

Two points have to be considered to complete the online learning. As we don't expect the noise type to change rapidly, we can reduce the computational complexity by updating the noise dictionary less frequently. Also, as an inherent property of NMF, good initializations can improve the dictionary learning. To address these two issues, we use a simple approach based on a sliding window concept. Let us define a local buffer $\underline{\mathbf{m}} \in \mathbb{R}_+^{K \times N_2}$ that stores the last N_2 observed noisy DFT magnitudes. Every time we observe a new frame, the columns in $\underline{\mathbf{m}}$ are shifted to the left and the most recent frame is stored at the rightmost column. When the local buffer is full, i.e., N_2 new frames have been observed, a number of frames (let's say q frames) that have the lowest energies are chosen to update the main buffer $\underline{\mathbf{n}}$. Note that to do this we don't use any voice

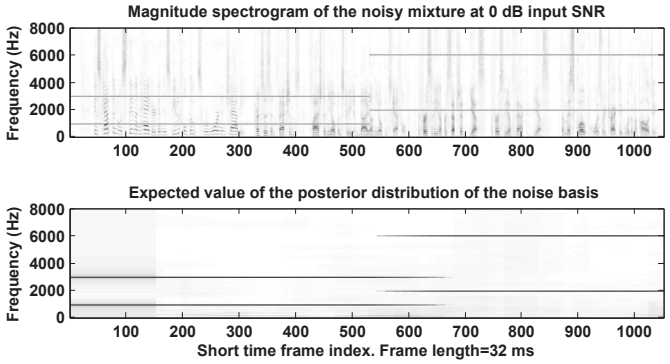

Fig. 4. Demonstration of the noise basis adaptation. The top panel shows a mixture magnitude spectrogram in which a sinusoidal noise signal (having two harmonics corresponding to the horizontal lines) is added to a speech signal at 0 dB input SNR. The bottom panel depicts a single noise basis vector over time that is adapted using the noisy mixture. See the text for more explanation.

activity detector. Hence, the columns in $\underline{\mathbf{n}}$ are shifted to the left and new data is stored on the rightmost columns of the buffer. We now apply the KL-NMF on $\underline{\mathbf{n}}$ for few iterations, and use the obtained basis matrix to initialize the posterior means of the noise basis matrix. Then, the iterations of variational Bayes (using both speech and noise basis matrices) are continued until convergence.

One of the important parameters in our online learning is N_1, size of the main buffer. Although a large buffer reduces the overfitting risk, it slows down the adaption speed of the basis matrix. The latter causes the effect of the previous noise to fade out slowly, which will be illustrated in the following example. In our experiments, we set $N_1 = 50$, $N_2 = 15$, $q = 5$. Our approach of the basis adaption is independent of the underlying SNR.

Fig. 4 provides a demonstration of the online noise basis learning using a toy example. For this example, a noisy signal (at 0 dB SNR) is obtained by adding two different sinusoidal noise signals to the speech waveform at a sampling rate of 16 kHz. A frame length of 32 ms with 50% overlap and a Hann window was utilized to implement the DFT. We learned a single noise basis vector ($I^{(n)} = 1$) from the noisy mixture. As depicted in the lower panel of Fig. 4, the noise basis is adapted correctly to capture the changes in the noise spectrum. BNMF-based speech enhancement resulted to a 13 dB improvement in source to distortion ratio (SDR) [42] and a 0.9 MOS improvement in PESQ [43] for this example.

As Fig. 4 demonstrates, the proposed online learning has introduced a latency of around 15 frames in the adaption of the noise basis. In general, this delay depends on both N_2 and the time alignment of the signals, but it is always upper bounded by $2N_2 - q$ short-time frames. Moreover, Fig. 4 shows a side effect of the sliding window where the effect of the previous noise is fed out slowly (depending on the parameters N_1, N_2 and q). However, in a practical scenario, the effect of this latency and slow decay are not as clear as this toy example because the noise characteristics change gradually and not abruptly.

An additional approach to adapt the noise basis is to update the basis matrix in each short-time frame. In this view, variational Bayes is applied to each noisy frame to obtain the posterior distribution of both the NMF coefficients and the noise basis matrix. However, our simulations showed that this approach is not robust enough to changes in the noise type. In fact, to capture the noise spectrogram changes and at the same time not overfit to a single frame, a tradeoff has to be considered in constructing the priors for the noise dictionary, which was difficult to achieve in our simulations.

C. Informative Priors for NMF Coefficients

To apply variational Bayes to the noisy signal, we use the temporal dependencies of data to assign prior distributions for the NMF coefficients $\mathbf{V}$. Both BNMF-based methods from III-A and III-B use this approach to recursively update the prior distributions. To model temporal dependencies and also to account for the non-stationarity of the signals, we obtain a prior for $\mathbf{V}_t$ by widening the posterior distributions of $\mathbf{V}_{t-1}$. Recalling (6), let the state-conditional prior distributions be: $f_{V_{it}}(v_{it} \mid x_t) = \mathcal{G}(v_{it}; \phi_{it}[x_t], \theta_{it}[x_t]/\phi_{it}[x_t])$ where state dependency is made explicit through the notation $[x_t]$. For this gamma distribution we have:

$$E\left(V_{it} \mid x_t\right) = \theta_{it}\left[x_t\right], \quad \frac{\sqrt{\text{var}\left(V_{it} \mid x_t\right)}}{E\left(V_{it} \mid x_t\right)} = \frac{1}{\sqrt{\phi_{it}\left[x_t\right]}}, \tag{17}$$

where var$(\cdot)$ represents the variance. We assign the following recursively updated mean value to the prior distribution:

$$\theta_{it}\left[x_t\right] = \alpha\theta_{i,t-1}\left[x_t\right] + (1-\alpha)\, E\left(V_{i,t-1} \mid \mathbf{y}_{t-1}, x_t\right), \tag{18}$$

where the value of α controls the smoothing level to obtain the prior. Note that due to the recursive updating, θ_{it} is dependent on all the observed noisy data $\mathbf{y}_1^{t-1}$.

In (17), different shape parameters are used for the speech and noise NMF coefficients, but they are constant over time. Thus, $\phi_{it} = \phi_{i,t-1} = \ldots \phi_{i1}$, also $\phi_{it} = \phi^{(s)}$ for $i = 1, \ldots I^{(s)}$, and $\phi_{it} = \phi^{(n)}$ for $i = I^{(s)}+1, \ldots I^{(s)} + I^{(n)}$. Moreover, different noise types are allowed to have different shape parameters. In this form of prior, the ratio between the standard deviation and the expected value is the same for all the NMF coefficients for a source signal. The shape parameter ϕ represents the uncertainty of the prior which in turn corresponds to the non-stationarity of the signal being processed. We can learn this parameter in the training stage using the clean speech or noise signals. Hence, at the end of the training stage, the shape parameters of the posterior distributions of all the NMF coefficients are calculated and their mean value is taken for this purpose. Using this approach for the speech signal results in $\phi^{(s)} = 3 \sim 5$. However, the noise reduction simulations suggest that having an uninformative prior for speech (a small value for $\phi^{(s)}$) leads to a better performance unless the noise signal is more non-stationary than the speech signal, e.g., keyboard or machine gun noises. Therefore, in our experiments we used a relatively flat prior for the speech NMF coefficients ($\phi^{(s)} \ll 1$) that gives the speech BNMF model greater flexibility.

Our experiments show that the optimal amount of smoothing in (18) depends on the long-term SNR (or global SNR). For

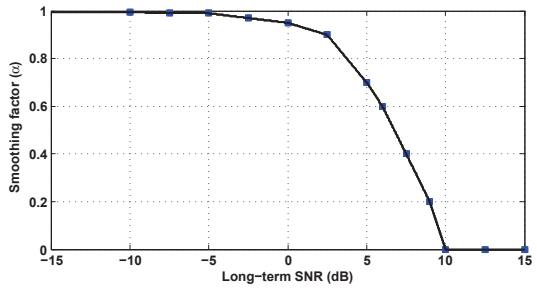

Fig. 5. An empirical α-SNR curve, which is used in our experiments. The figure shows that for low input SNRs (high noise levels) a high degree of smoothing should be applied to update the mean values of the prior distributions for NMF coefficients (18), and vice versa.

low SNRs (high level of noise) a strong smoothing ($\alpha \to 1$) improves the performance by reducing unwanted fluctuations while for high SNRs a milder smoothing ($\alpha \to 0$) is preferred. The latter case corresponds to obtaining the mean value θ directly using the information from the previous time frame. Here, in contrast to [31], we use an SNR-dependent value for the smoothing factor. Fig. 5 shows an $\alpha -$ SNR curve that we obtained using computer simulations and was used in our experiments.

To calculate the long-term SNR from the noisy data, we implemented the approach proposed in [44] that works well enough for our purpose. This approach assumes that the amplitude of the speech waveform is gamma-distributed with a shape parameters fixed at 0.4, and that the background noise is Gaussian-distributed, and that speech and noise are independent. Under these assumptions, authors have modeled the amplitude of the noisy waveform with a gamma distribution and have shown that the maximum likelihood estimate of the shape parameter is uniquely determined from the long-term SNR [44].

IV. Experiments and Results

We evaluate and compare the proposed NMF-based speech enhancement systems in this section. The experiments are categorized as supervised and unsupervised speech enhancement methods. In Subsection IV-A, we evaluate the noise reduction systems where for each noise type we have access to some training data. Evaluation of the unsupervised denoising schemes is presented in IV-B, where we assume that we don't have training data for some of the noise types.

In our simulations, all the signals were down-sampled to 16 kHz and the DFT was implemented using a frame length of 512 samples and 0.5-overlapped Hann windows. The core test set of the TIMIT database (192 sentences) [45] was exploited for the noise reduction evaluation. The signal synthesis was performed using the overlap-and-add procedure. SNR was

For all the BNMF-based methods, a universal speaker-independent speech model with 60 basis vectors is learned using the training data from the TIMIT database. The choice of dictionary size is motivated by our previous study [46]. Moreover, for the BNMF-based approaches the long-term SNR was estimated using [44] and we used Fig. 5 to apply an SNR-dependent smoothing to obtain the priors.

As reviewed in Section II, the method introduced in [30] factorizes the whole spectrogram of the noisy signal, and therefore, is not causal. In order to make it more practical, we considered two causal extensions of this work and evaluated their performance in this section. The first extension is a supervised approach that works frame by frame. Here, we trained one universal NHMM (100 states and 10 basis vectors per state) for speech and one single-state NHMM for each noise type. To achieve causality, we simply replaced the forward-backward algorithm with the forward algorithm in which the NMF coefficients from the previous timestamp were used to initialize the current ones. As the other extension, we adapted an online noise dictionary learning, similarly to Section III-B.

A. Noise Reduction Using a-Priori Learned NMF Models

We evaluated five variants of NMF-based enhancement methods for three noise types. The considered noise types include factory and babble noises from the NOISEX-92 database [47] and city traffic noise from Sound Ideas [48]. Although all of these three noises are considered non-stationary, the city traffic noise is very non-stationary since it includes mainly horn sounds. We implemented five NMF-based algorithms including:

1) BNMF-HMM: we used (10) in which the noise-type is not known in advance.
2) General-model BNMF: we trained a single noise dictionary by applying BNMF on a long signal obtained by concatenating the training data of all three noises. For the enhancement, (15) was used regardless of the underlying noise type.
3) Oracle BNMF: this is similar to BNMF-HMM but the only difference is that instead of the proposed classifier an oracle classifier is used to choose a noise model for enhancement, i.e., the noise type is assumed to be known a priori and its offline-learned basis matrix is used to enhance the noisy signal. Therefore, this approach is an ideal case of BNMF-HMM.
4) Oracle ML: this supervised method is the maximum likelihood implementation of the Oracle BNMF in which KL-NMF in combination with (2) is used to enhance the noisy signal. Similar to the previous case, an oracle classifier is used to choose a noise model for enhancement. The term ML reflects the fact that KL-NMF arises as the maximum likelihood solution of (4) and (5).
5) Oracle NHMM: this is basically the supervised causal NHMM, as explained earlier in IV. Similar to cases (3) and (4), the noise type is assumed to be known in advance.

The number of basis vectors in the noise models were set using simulations performed on a small development set. For BNMF and KL-NMF methods, we trained 100 basis vectors for each noise type. Also, 200 basis vectors were learned for the general noise model. For NHMM, a single state with 100 basis vectors were learned for factory and city traffic noises while 30 basis vectors were pre-trained for babble noise since it provided a better performance.

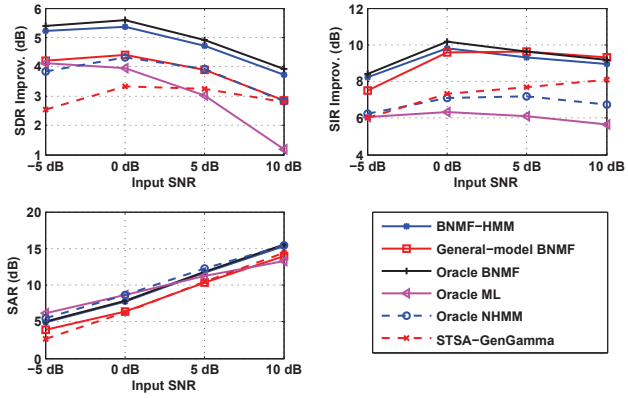

Fig. 6. BSS-Eval measures [42] to evaluate and compare the supervised NMF-based denoising algorithms. The BNMF-based schemes are described in Subsection III-A. Here, the prefix "Oracle" is used for the variants where the noise type is known a priori. The results are averaged over different noise types. For the SDR and SIR, improvements gained by the enhancement systems are shown.

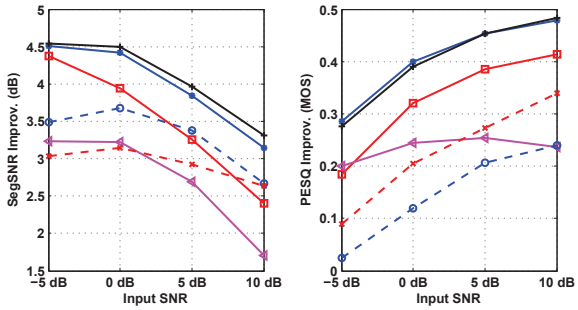

Fig. 7. PESQ and Segmental SNR (SegSNR) improvements gained by the supervised enhancement systems. Legend of this figure is similar to that of Fig. 6.

The performance of the NMF-based methods is compared to a speech short-time spectral amplitude estimator using super-Gaussian prior distributions [7], which is referred to as STSA-GenGamma. Here, we used [12] to track the noise PSD, and we set $\gamma = \nu = 1$ since it is shown to be one of the best alternatives [7]. This algorithm is considered in our simulations as a state-of-the-art benchmark to compare NMF-based systems.

Fig. 6 shows the source to distortion ratio (SDR), source to interference ratio (SIR), and source to artifact ratio (SAR) from the BSS-Eval toolbox [42]. SDR measures the overall quality of the enhanced speech while SIR and SAR are proportional to the amount of noise reduction and inverse of the speech distortion, accordingly. For SDR and SIR, the improvements gained by the noise reduction systems are shown. Several interesting conclusions can be drawn from this figure.

The simulations show that the Oracle BNMF has led to the best performance, which is closely followed by BNMF-HMM. The performance of these two systems is quite close with respect to all three measures. This shows the superiority of the BNMF approach, and also, it indicates that the HMM-based classification scheme is working successfully. Another interesting result is that except for the Oracle ML, the other NMF-based techniques outperform STSA-GenGamma. The ML-NMF approach gives a poor noise reduction particularly at high input SNRs. These results were confirmed by our informal listening tests.

Moreover, the figure shows that the Oracle NHMM and General-model BNMF methods lead to similar SDR values. However, these two methods process the noisy signal differently. The NHMM method doesn't suppress a lot of noise but it doesn't distort the speech signal either (i.e., SAR is high). This is reversed for the General-model BNMF. Furthermore, comparing BNMF-HMM and General-model BNMF confirms an already reported observation [14], [16] that using many small noise-dependent models is superior to a large noise-independent model.

Fig. 7 provides the experimental results using segmental SNR (SegSNR) [49, ch. 10], which is limited to the range $[-10\text{dB}, 30\text{dB}]$, and perceptual evaluation of speech quality (PESQ) [43]. As it can be seen in the figure, the BNMF-based methods have led to the highest SegSNR and PESQ improvements. These results verify again the excellence of the BNMF strategies. Moreover, it is interesting to note that the NHMM method has not been very successful in improving the quality of the noisy speech with respect to the PESQ measure.

To study specifically the classification part of the BNMF-HMM algorithm, we analyzed the output of the classifier. Fig. 8 provides the result of this experiment. To have a clearer representation, the probability of each noise type in (16) is smoothed over time and is depicted in Fig. 8. Here, the classifier is applied to a noisy signal at 0 dB input SNR. The underlying noise type is given as the titles of the subplots. As it can be seen in the figure, the classifier works reasonably well in general. Most of the wrong classifications correspond to the case where the true noise type is confused with the babble noise. One reason for this confusion is due to the nature of babble noise. If the short-time spectral properties of the noise are not very different from those of babble, the union of speech and babble basis vectors can explain any noisy signal by providing a very good fit to the speech part. However, as shown in Fig. 6 and Fig. 7, this confusion has reduced the performance only very marginally.

B. Experiments with Unsupervised Noise Reduction

This subsection is devoted to investigating the performance of the unsupervised NMF-based speech enhancement systems. For this purpose, we considered 6 different noise types including factory and babble noises from the NOISEX-92 database [47], and city traffic, highway traffic, ocean, and hammer noises from Sound Ideas [48]. Among these, ocean noise can be seen as a stationary signal in which the noise level changes up to ± 20 dB. All the signals were concatenated before processing.

We evaluated three NMF-based enhancement systems using a general speech model, which is learned similarly to Subsection IV-A. We considered Online BNMF (Subsection III-B) and Online NHMM (as explained earlier in Section IV). Additionally, we included the BNMF-HMM in the comparison. The considered BNMF-HMM model was identical to that of

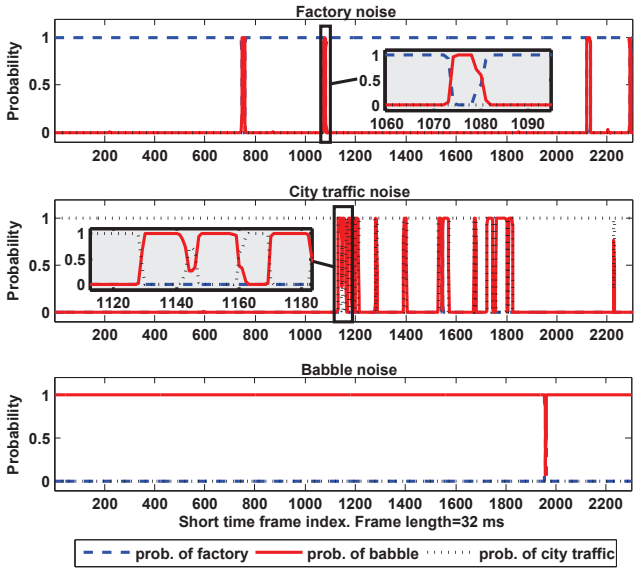

Fig. 8. Result of the noise classifier where (16) is smoothed over time and is plotted for a mixture at 0 dB input SNR. The underlying noise type is given in the titles of the subplots (which corresponds to factory, city traffic, and babble noises, respectively, from top to bottom). In each subplot, the probability of three noise classes (factory, city traffic, and babble noises) are shown. For visibility, two small segments are magnified and shown in the figure.

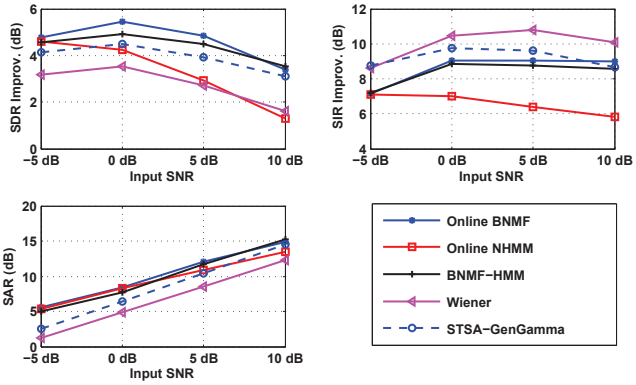

Fig. 9. SDR and SIR improvements and SAR measure [42] to evaluate and compare the unsupervised NMF-based denoising algorithms. For the Online BNMF and Online NHMM variants, the noise basis matrix is learned online from the noisy data, explained in III-B. The results are averaged over different noise types. For the BNMF-HMM approach, similar to Fig. 6, only three noise models are learned.

Subsection IV-A, i.e., we learned only three models for factory, babble and city traffic noises. For the other noise types, the method is allowed to use any of these models to enhance the noisy signal according to (10). Furthermore, we included two state-of-the-art approaches in our experiments: The STSA-GenGamma approach, identical to that of Subsection IV-A, and a Wiener filter in which the noise PSD was estimated using [12] and a decision-directed approach [50] was used to implement the filter. Here, the final gain applied to the noisy signal was limited to be larger than 0.1, for perceptual reasons [51].

For the online BNMF and online NHMM algorithms, we learned $I^{(n)} = 30$ basis vectors for noise. Learning a large basis matrix is this case can lead to overfitting since the dictionary is adapted given a small number of observations ($N_1 = 50$ in our experiments). This was also verified in our computer simulations. Hence, in contrast to the supervised methods for which we learned 100 basis vectors for each noise, we learned a smaller dictionary for online algorithms.

Fig. 9 shows the objective measures from BSS-Eval [42] for different algorithms. As it can be seen in the figure, Online BNMF has outperformed all the other systems. This method introduces the least distortion in the enhanced speech signal while performing moderate noise reduction. On the other hand, Wiener filter and STSA-GenGamma reduce the interfering noise greatly with the cost of introducing artifacts in the output signal.

Online NHMM outperforms the Wiener and STSA-GenGamma algorithms at low input SNRs with respect to SDR but for high input SNRs the performance of the algorithm is the worst among all the competing methods. Also, the amount of noise suppression using Online NHMM is the least among different methods.

Moreover, Fig. 9 shows that STSA-GenGamma provides a higher-quality enhanced speech signal than the Wiener filter. This is reported frequently in the literature, e.g. [7].

Another interesting result that can be seen in Fig. 9 is that Online BNMF outperforms the BNMF-HMM. The difference in the performance is even larger with respect to SegSNR and PESQ, shown in Fig. 10. As it is shown in this figure, Online BNMF outperforms the BNMF-HMM (and the other methods) with a large margin.

To have a better understanding on how Online BNMF and BNMF-HMM schemes behave for different noise types, we evaluated SDR and PESQ over short intervals of time. To do so, the noisy and enhanced speech signals were windowed into segments of 5 seconds and then for each segment a SDR and PESQ value was calculated. Fig. 11 shows such results as a function of window index. The boundary of the underlying noise types is shown in green in six different levels in which segments belong to factory, babble, city traffic, highway traffic, ocean, and hammer noises, respectively from left to right. As can be seen in the figure, for the first three noise types for which a noise-dependent BNMF model is learned offline the BNMF-HMM approach works marginally better than the Online BNMF. But, for the last three noise types Online BNMF outperforms BNMF-HMM significantly. The difference is highest for the hammer noise; this is due to our observation that the hammer noise differs more from either factory, babble or city traffic noises than highway traffic or ocean noises do. Therefore, neither of the pre-trained models can explain the hammer noise well, and as a result, the overall performance of the BNMF-HMM degrades whenever there is a large mismatch between the training and the testing signals.

A final remark about the Online BNMF and BNMF-HMM can be made considering the computational complexity. In our simulations (where we didn't use parallel processing techniques), Online BNMF runs twice as fast as BNMF-HMM with three states. Moreover, our Matlab implementation of the Online BNMF runs in approximately 5-times real time in a PC with 3.8 GHz Intel CPU and 2 GB RAM.

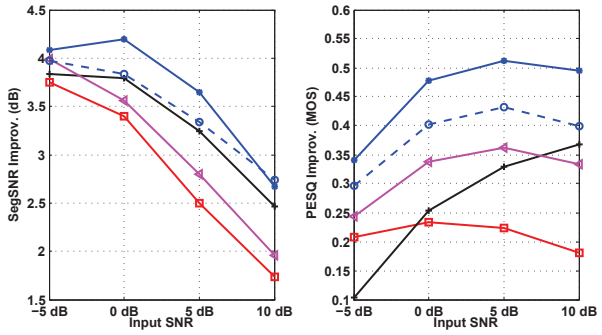

Fig. 10. PESQ and Segmental SNR (SegSNR) improvements gained by the unsupervised enhancement systems. Legend of this figure is similar to that of Fig. 9.

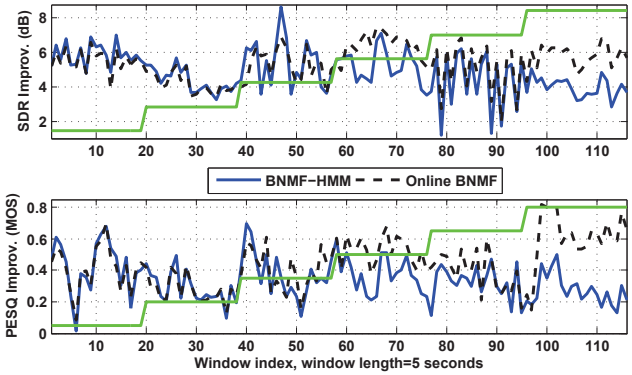

Fig. 11. SDR and PESQ measured over short intervals of 5-second long. Six different levels shown in green correspond to factory, babble, city traffic, highway traffic, ocean, and hammer noises, respectively from left to right. For the BNMF-HMM approach, only three noise models corresponding to the first three noises are learned; for the other noise types, the estimator chooses a model that can describe the noisy observation better than the other models.

V. Conclusions

This paper investigated the application of NMF in speech enhancement systems. We developed speech enhancement methods using a Bayesian formulation of NMF (BNMF). We proposed two BNMF-based systems to enhance the noisy signal in which the noise type is not known a priori. We developed an HMM in which the output distributions are assumed to be BNMF (BNMF-HMM). The developed method performs a simultaneous noise classification and speech enhancement and therefore doesn't require the noise type in advance. Another unsupervised system was constructed by learning the noise BNMF model online, and is referred to as Online BNMF.

Our experiments showed that a noise reduction system using a maximum likelihood (ML) version of NMF—with a universal speaker-independent speech model—-doesn't outperform state-of-the-art approaches. However, by incorporating the temporal dependencies in form of prior distributions and using optimal MMSE filters, the performance of the NMF-based methods increased considerably. The Online BNMF method is faster than the BNMF-HMM and was shown to be superior when the underlying noise type was not included in the training data. Our simulations showed that the suggested systems outperform the Wiener filter and an MMSE estimator of speech short-time spectral amplitude (STSA) using super-Gaussian priors with a high margin while they are not restricted to know any priori information that is difficult to obtain in practice.

Acknowledgment

The authors are grateful to Gautham J. Mysore for providing a Matlab implementation of the NHMM approach in [30].

References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, apr. 1979.

[2] J. S. Lim and V. O. Alan, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, dec. 1979.

[3] V. Grancharov and J. S. B. Kleijn, "On causal algorithms for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 3, pp. 764–773, may 2006.

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

[5] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 13, no. 5, pp. 845–856, sep. 2005.

[6] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models," *Signal Process.*, vol. 86, no. 4, pp. 698–709, apr. 2006.

[7] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum Mean-Square Error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1741–1752, aug. 2007.

[8] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Communication*, vol. 49, no. 2, pp. 134–143, feb. 2007.

[9] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, sep. 2012.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, jul. 2001.

[11] I. Cohen, "Noise spectrum estimation in adverse environments : Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, sep. 2003.

[12] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, mar. 2010, pp. 4266–4269.

[13] T. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, sep. 1996.

[14] S. Srinivasan, J. Samuelsson, and W. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 163–176, jan. 2006.

[15] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, apr. 1992.

[16] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 445–455, sep. 1998.

[17] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 882–892, mar. 2007.

[18] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," *IEEE Signal Process. Letters*, vol. 20, no. 3, pp. 253–256, mar. 2013.

[19] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain," *Speech Communication*, vol. 55, no. 2, pp. 205–220, feb. 2013.

[20] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame level noise classification in mobile environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, mar. 1999, pp. 237–240.
[21] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Process. Systems Conf. (NIPS)*, 2000, pp. 556–562.
[22] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York: John Wiley & Sons, 2009.
[23] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 1–12, jan. 2007.
[24] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.
[25] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.
[26] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, may 2013.
[27] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 411–414.
[28] M. Schmidt and J. Larsen, "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *IEEE Workshop on Machine Learning for Signal Process. (MLSP)*, oct. 2008, pp. 486–491.
[29] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, 2011, pp. 45–48.
[30] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, may. 2011, pp. 17–20.
[31] N. Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using Bayesian NMF with recursive temporal updates of prior distributions," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2012, pp. 4561–4564.
[32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.
[33] A. Lefevre, F. Bach, and C. Févotte, "Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoustics (WASPAA)*, 2011, pp. 313–316.
[34] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009, article ID 785152, 17 pages.
[35] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Advances in models for acoustic processing workshop, NIPS*, 2006.
[36] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.
[37] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new approach for speech enhancement based on a constrained nonnegative matrix factorization," in *IEEE Int. Symp. on Intelligent Signal Process. and Communication Systems (ISPACS)*, dec. 2011, pp. 1–5.
[38] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, 2013.
[39] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2011, pp. 1217–1220.
[40] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.
[41] J. A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," U.C. Berkeley, Tech. Rep. ICSI-TR-97-021, 1997.
[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.
[43] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.
[44] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 2598–2601.
[45] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus." Philadelphia: Linguistic Data Consortium, 1993.
[46] N. Mohammadiha and A. Leijon, "Model order selection for non-negative matrix factorization with application to speech enhancement," KTH Royal Institute of Technology, Tech. Rep., 2011.
[47] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, jul. 1993.
[48] B. Nimens *et al.*, "Sound ideas: sound effects collection," ser. 6000, http://www.sound-ideas.com/6000.html.
[49] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. CRC Press, 2007, vol. 30.
[50] Y. Ephraim and I. Cohen, *Recent Advancements in Speech Enhancement*. in The Electrical Engineering Handbook, CRC Press, 2005.
[51] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 2, mar. 1999, pp. 789–792.

Nasser Mohammadiha (S'11) received the M.Sc. degree in electronics engineering from Sharif University of Technology, Tehran, Iran, in 2006. He worked on digital hardware and software design until 2008.

He is currently pursuing a Ph.D. degree in telecommunications at the Department of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include speech and image processing, mainly speech enhancement, machine learning applied to audio, and statistical signal modeling. He is a student member of the IEEE.

Paris Smaragdis (M'03) is faculty in the Computer Science and the Electrical and Computer Science departments at the University of illinois at Urbana-Champaign. He completed his graduate and postdoctoral studies at MIT, where he conducted research on computational perception and audio processing. Prior to the University of Illinois he was a senior research scientist at Adobe Systems and a research scientist at Mitsubishi Electric Research Labs, during which time he was selected by the MIT Technology Review as one of the top 35 young innovators of 2006. Paris' research interests lie in the intersection of machine learning and signal processing, especially as they apply to audio problems.

Arne Leijon (M'10) received the MS degree in engineering physics in 1971, and the Ph.D. degree in information theory in 1989, both from Chalmers University of Technology, Gothenburg, Sweden.

He has been a professor of hearing technology at the Sound and Image Processing (SIP) Laboratory at the KTH Royal Institute of Technology, Stockholm, Sweden, since 1994. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these aids, based on psychoacoustic modeling of sensory information transmission and subjective sound quality. He is a member of the IEEE.