http://kth.diva-portal.org

This is an author produced version of a paper presented at the *IEEE International Workshop on Machine Learning for Signal Processing (MLSP), September 22-25, 2013* .This paper has been peer-reviewed but does not include the final publisher proof-corrections or proceedings pagination.

© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Citation for the published paper:

Nasser Mohammadiha, Paris Smaragdis and Arne Leijon

Simultaneous Noise Classification and Reduction Using a Priori Learned Models

Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP).

Access to the published version may require subscription.

Published with permission from: IEEE

SIMULTANEOUS NOISE CLASSIFICATION AND REDUCTION USING A PRIORI LEARNED MODELS

Nasser Mohammadiha[†] *Paris Smaragdis*[‡] *Arne Leijon*[†]

[†] KTH Royal Institute of Technology
Dept. of Electrical Engineering, Stockholm, Sweden
[‡]University of Illinois at Urbana-Champaign
Dept. of Computer Science and Dept. of Electrical and Computer Engineering
Adobe Systems Inc.

ABSTRACT

Classifying the acoustic environment is an essential part of a practical supervised source separation algorithm where a model is trained for each source offline. In this paper, we present a classification scheme that is combined with a probabilistic nonnegative matrix factorization (NMF) based speech denoising algorithm. We model the acoustic environment with a hidden Markov model (HMM) whose emission distributions are assumed to be of NMF type. We derive a minimum mean square error (MMSE) estimator of clean speech signal in which the state-dependent speech estimators are weighted according to the state posterior probabilities (or probabilities of different noise environments) and are summed. Our experiments show that the proposed method outperforms state-of-the-art substantially and that its performance is very close to an oracle case where the noise type is known in advance.

Index Terms— Nonnegative matrix factorization, acoustic environment classification, supervised speech enhancement

1. INTRODUCTION

In this paper, we develop a single-channel speech enhancement system in which a supervised noise reduction approach is combined with an environmental noise classification technique. Such a solution is of interest for different applications such as hearing aids and speech communication over telephone and internet.

In general, speech enhancement methods can be categorized into two broad classes of unsupervised and supervised approaches. In the unsupervised methods, e.g., Wiener filtering and methods based on super-Gaussian prior distributions for speech DFT coefficients [1], estimation of the clean speech signal is carried out without any prior information on the noise type. For the supervised methods, on the other hand, speech and noise models are trained offline using some training data. Then, assuming that the noise type is known, the noise reduction task is performed. Some examples of this class of algorithms include hidden Markov model (HMM) based methods [2] and NMF-based techniques [3, 4].

The main difficulty of the unsupervised methods is the noise power spectral density (PSD) estimation, e.g. [5], which is a challenging task for non-stationary background noises. One advantage of the supervised methods is that there is no need to estimate the noise PSD using a separate algorithm. Moreover, the supervised approaches have been shown to produce better results compared to the unsupervised methods [2, 6, 7].

The main practical issue of the supervised approaches is that the required prior information on the noise type and speaker identity has to be estimated online since this information is usually not available in an online system. This task can be solved using a built-in classification scheme [2], or using an external environmental noise classification algorithm, e.g., [8]. In this paper we propose a probabilistic built-in classification technique that is integrated into a speech enhancement system.

NMF is a technique to approximate a nonnegative matrix $\mathbf{Y}$ as a product of a basis matrix $\mathbf{B}$ and an activation matrix $\mathbf{V}$, i.e., $\mathbf{Y} \approx \mathbf{BV}$ [9]. In speech processing, $\mathbf{Y}$ is usually the spectrogram of the speech signal. NMF has been recently used to estimate the clean speech from a noisy observation [3,4,6,10]. A main focus of most of these approaches is to use the temporal dynamics in an NMF-based speech denoising or separation algorithm [4, 6, 11].

When applied to the speech source separation, a good separation can be expected only when speaker-dependent basis matrices are learned. However, for the noise reduction even if a universal speaker-independent basis matrix of speech is learned a good enhancement can be achieved [4]. In some cases when the interfering noise exhibits speech-like properties, e.g. babble noise, to have a better enhancement additional constraints should be imposed into NMF. For example, assuming that babble waveform is obtained as a sum of different speech signals a nonnegative HMM is proposed in [7] to model the babble noise in which babble basis is identical to the speech basis. The obtained babble model is then used to devise a noise reduction method.

The available NMF-based noise reduction systems suffer from the problem that they either need some a priori information that is difficult to obtain in practice or they are not appropriate for online applications. For example, the approach presented in [3] is only applicable in batch mode, meaning that the whole noisy signal is assumed to be observed for the purpose of enhancement. Likewise, the semi-supervised approach proposed in [10], although it does not need to know the noise type in advance, requires the entire spectrogram of the noisy signal to denoise the observed signal.

Methods proposed in [4, 6, 12] require to know the noise type in advance to use noise-dependent basis matrices to enhance the noisy signal. For some applications where user is allowed to choose a scenario from a given set, this assumption can be feasible. However, in general, a separate environmental noise classifier or a noise-independent basis matrix have to be used in these approaches to be used in practice.

In this paper, we further develop our proposed method in [4] to

design a noise reduction system in which the noise-type is not known in advance. In the proposed method, the temporal dependencies are used to construct informative prior distributions to be applied in a Bayesian formulation of NMF (BNMF). In contrast to [4], the prior distributions are signal to noise ratio (SNR) dependent where SNR is estimated online from the noisy mixture. Then, we develop an HMM with output density functions given by the BNMF to design a simultaneous noise classification and reduction system. We derive a minimum mean square error (MMSE) estimator for the speech signal in which the noise type does not need to be known a priori. In the proposed scheme, the classification is done using the noisy input and is not restricted to be applied at only the speech pauses as it is in [2]. We evaluate the proposed system in different noise levels and compare its performance to state-of-the-art and show that it is clearly superior to the competing methods.

2. SIMULTANEOUS NOISE CLASSIFICATION AND REDUCTION USING BNMF-HMM

We explain our proposed method to perform simultaneous noise classification and suppression in this section. We use an HMM to model the acoustic environment where each state of the HMM corresponds to a noise type. Given the hidden state, we model the noise magnitude spectrogram using a probabilistic NMF that will be explained in this section. Also, all the states share some common parameters that include an NMF model for the speech magnitude spectrogram and an estimate of the long-term signal to noise ratio (SNR). Having this setup, we derive an MMSE estimator for the clean speech signal that is conditioned on the observed noisy signal.

2.1. A Probabilistic Nonnegative Factorization

Compared to the deterministic formulations of NMF, e.g. [9], a probabilistic NMF provides an easier way to incorporate our prior knowledge about the sources. Moreover, we can derive optimal estimates of the desired signal, e.g., MMSE estimate, in a probabilistic framework. Therefore, in this paper we assume that the speech, noise, and noisy magnitude spectrograms are random variables. Each time-frequency bin of a spectrogram is assumed to be a sum of I latent variables as:

$$y_{kt} \quad = \quad \sum_{i=1}^{I} z_{kit}, \tag{1}$$

where k and t denote the frequency and time indices, respectively, and each hidden variable z is assumed to be drawn from a Poisson distribution whose mean value is given by $b_{ki}v_{it}$, i.e.,

$$f\left(z_{kit}\right) \quad = \quad \left(b_{ki}v_{it}\right)^{z_{kit}} e^{-b_{ki}v_{it}} / \left(z_{kit}!\right), \tag{2}$$

where $z!$ is the factorial of z. Using the properties of the Poisson distribution we can see that y_{kt} is also drawn from a Poisson distribution whose mean value is given by $\sum_i b_{ki}v_{it}$. We use this mean value to provide an NMF approximation of the observed data. Writing in matrix form we have $\mathbf{Y} \approx \mathbf{BV}$ where $\mathbf{Y} = [y_{kt}]$, $\mathbf{B} = [b_{ki}]$, and $\mathbf{V} = [v_{it}]$. The maximum likelihood (ML) estimates of the parameters $\mathbf{B}$ and $\mathbf{V}$ can be obtained using an EM algorithm [13], and the result is identical to the well-celebrated multiplicative update rules for NMF using Kullback-Leibler (KL-NMF) divergence [9].

To provide a way to impose our prior knowledge into the factorization, the nonnegative factors are further assumed to be stochastic. We assume that each element of the basis matrix $\mathbf{B}$ and activation matrix $\mathbf{V}$ are drawn from a gamma distribution as follows:

$$\begin{aligned} f\left(v_{it}\right) &= \mathcal{G}\left(v_{it}; \phi_{it}, \theta_{it}/\phi_{it}\right), \\ f\left(b_{ki}\right) &= \mathcal{G}\left(b_{ki}; \psi_{ki}, \gamma_{ki}/\psi_{ki}\right), \end{aligned} \tag{3}$$

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 (thus mean value is given by $\phi\theta$), and $\Gamma\left(\phi\right)$ is the gamma function. ϕ, θ, ψ and γ are referred to as the hyperparameters.

Next, we need to infer the posterior distribution of the variables. As the exact Bayesian inference is intractable for (1), (2), and (3) a variational Bayes approach has been proposed in [13] to obtain the approximate posterior distributions of the variables. Hence, in an iterative scheme the current parameters of the posterior distributions of $\mathbf{Z}$ are used to update the parameters of the posterior distributions of $\mathbf{B}$ and $\mathbf{V}$, and these new parameters are used to update the posterior distributions of $\mathbf{Z}$ in the next iteration. The iterations are carried on until convergence. The posterior distributions for $\mathbf{z}_{k,:,t}$ 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 [13]. 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).

2.2. BNMF-HMM Structure

In this section, we describe the proposed BNMF-HMM method. Let us consider M noise types for which we are able to gather some training data. We first use some appropriate training data to obtain a BNMF model for each type of the considered noises and for the speech signal offline. Here, we consider a general speaker-independent model of the speech signal, which does not introduce any limitation in the approach.

The structure of the BNMF-HMM is shown in Fig. 1. Each state of the HMM has some state-dependent parameters, which are just 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 SNR that will be used for the enhancement. To complete the Markovian model, we predefine a state transition matrix whose diagonal elements are set to some high values, and the rest of its elements are set to some small values such that each row of the transition matrix sums to one. Moreover, each element of the initial state probability is also set to $1/M$.

We model the magnitude spectrogram of the clean speech and noise signals by (1). To train a BNMF model, we need to obtain the posterior distribution of the basis matrix. During the training, we assign some sparse and broad prior distributions to $\mathbf{B}$ and $\mathbf{V}$ according to (3). After convergence of the variational Bayes approach, the posterior distributions of the noise basis matrix ($\mathbf{B}^{(n)}$) and speech basis matrix ($\mathbf{B}^{(s)}$) are stored to be used for the enhancement.

Let us denote the hidden state variable at time t by x_t that can take one of the possible outcomes $x_t = 1, 2, \ldots M$. Also, we show the magnitude of the discrete Fourier transform (DFT) coefficients of the speech, noise, and noisy signals by $\mathbf{S} = [s_{kt}]$,$\mathbf{N} = [n_{kt}]$ and $\mathbf{Y} = [y_{kt}]$, respectively. The vector of noisy DFT magnitudes, given the state x_t, is approximated as $\mathbf{y}_t = \mathbf{s}_t + \mathbf{n}_t$. To obtain the state-dependent distribution of $\mathbf{y}_t$, the parameters of the speech and noise basis parameters ($\mathbf{B}^{(s)}$, $\mathbf{B}^{(n)}$) are concatenated to get the parameters of the noisy basis matrix $\mathbf{B}$, i.e., $\mathbf{B} = [\mathbf{B}^{(s)}\ \mathbf{B}^{(n)}]$. Now, Eq. (1) is

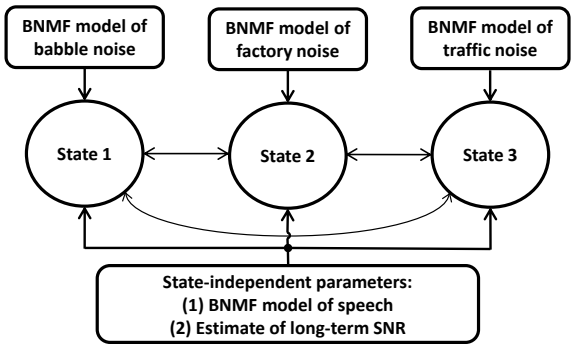

Fig. 1. A block diagram representation of BNMF-HMM with three states.

written as

$$y_{kt} = s_{kt} + n_{kt} = \sum_{i=1}^{I^{(s)}} z_{kit}^{(s)} + \sum_{i=1}^{I^{(n)}} z_{kit}^{(n)} = \sum_{i=1}^{I^{(s)}+I^{(n)}} z_{kit}, \quad (4)$$

where $I^{(s)}$and $I^{(n)}$ are the number of the speech and noise basis vectors, respectively. As a result of (4), the distribution of the noisy DFT magnitudes is given by

$$f\left(y_{kt} \mid x_t, \mathbf{B}, \mathbf{v}_t\right) = \frac{\lambda_{kt}^{y_{kt}} e^{-\lambda_{kt}}}{y_{kt}!}, \quad (5)$$

where $\lambda_{kt} = \sum_i b_{ki} v_{it}$. Note that to keep the notations uncluttered in (5) we skip writing the state-dependency of $\mathbf{B}$ explicitly. The state-conditional likelihood of the noisy signal can now be computed by integrating over $\mathbf{B}$ and $\mathbf{v}_t$, as:

$$f\left(y_{kt} \mid x_t\right) = \int\int f\left(y_{kt} \mid \mathbf{B}, \mathbf{v}_t, x_t\right) f\left(\mathbf{B}, \mathbf{v}_t \mid x_t\right) d\mathbf{B} d\mathbf{v}_t. \quad (6)$$

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

$$f\left(\mathbf{y}_t \mid x_t\right) = \prod_k f\left(y_{kt} \mid x_t\right). \quad (7)$$

2.3. Speech Enhancement Using BNMF-HMM

As the first step of the enhancement, variational Bayes approach is applied to approximate the posterior distributions of the activation vector $\mathbf{v}_t$ by maximizing the variational lower bound on (7). 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 later.

The MMSE estimate of the speech DFT magnitudes can be shown to be [7]:

$$\hat{s}_{kt} = E\left(\mathsf{s}_{kt} \mid \mathbf{y}_t\right) = \frac{\sum_{x_t=1}^{M} \xi_t\left(\mathbf{y}_t, x_t\right) E\left(\mathsf{s}_{kt} \mid x_t, \mathbf{y}_t\right)}{\sum_{x_t=1}^{M} \xi_t\left(\mathbf{y}_t, x_t\right)}, \quad (8)$$

where $\xi_t\left(\mathbf{y}_t, x_t\right) = f(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1}) = f(\mathbf{y}_t \mid x_t) f(x_t \mid \mathbf{y}_1^{t-1})$, $\mathbf{y}_1^{t-1} = \{\mathbf{y}_1, \ldots \mathbf{y}_{t-1}\}$. Here, $f(x_t \mid \mathbf{y}_1^{t-1})$ is computed using the forward algorithm [14]. Since (6) can not be evaluated analytically, we approximate it by evaluating the integral at the mean value of the posterior distributions of $\mathbf{B}$ and $\mathbf{v}_t$, which are denoted by $\mathbf{B}'$ and $\mathbf{v}_t'$, respectively, as $f\left(y_{kt} \mid x_t\right) \approx f\left(y_{kt} \mid \mathbf{B}', \mathbf{v}_t', x_t\right)$. The state-dependent MMSE estimate of the speech DFT magnitudes $E(\mathsf{s}_{kt} \mid x_t, \mathbf{y}_t)$ in (8) can be obtained using the variational Bayes approach and is given by [4]:

$$E\left(\mathsf{s}_{kt} \mid x_t, \mathbf{y}_t\right) = \frac{\sum_{i=1}^{I^{(s)}} e^{E(\log \mathsf{b}_{ki} + \log \mathsf{v}_{it} \mid x_t, \mathbf{y}_t)}}{\sum_{i=1}^{I^{(s)}+I^{(n)}} e^{E(\log \mathsf{b}_{ki} + \log \mathsf{v}_{it} \mid x_t, \mathbf{y}_t)}} y_{kt}. \quad (9)$$

The time-domain enhanced speech signal is reconstructed using the noisy phase information. When the posterior distributions of the basis and activation matrices are very sharp (which happen for very large shape parameters in the gamma distribution), Eq. (9) takes a simple form of a Wiener filtering[1].

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

$$f\left(x_t \mid \mathbf{y}_1^t\right) = \frac{f(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1})}{\sum_{x_t} f(\mathbf{y}_t, x_t \mid \mathbf{y}_1^{t-1})}. \quad (10)$$

$f\left(x_t \mid \mathbf{y}_1^t\right)$ is the probability of each noise type given all the past noisy observations. Another likelihood-based classifications has been used in [2] for HMM-based denoising systems. Here, 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. In contrast, our proposed classification in (10) does not need any voice activity detector.

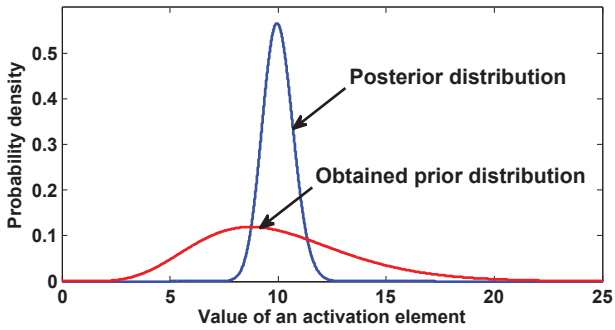

Fig. 2. The posterior distribution of the activations ($\mathbf{V}$) at past time frames are widened and are used as the prior distributions for the current time instance in a Bayesian framework. In this example, the mean value of the prior and posterior are the same while the variance of the prior is increased to reflect a higher uncertainty.

2.4. SNR-dependent Prior Distributions

To apply variational Bayes to the noisy signal, we can use the temporal dependencies of data to assign prior distributions to the activations $\mathbf{v}_t$. For this purpose 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 $\mathbf{v}_{t-1}$. Fig. 2 demonstrates such an approach using a toy example. Let the state-conditional prior distributions be: $f(v_{it} \mid x_t) = \mathcal{G}(v_{it}; \phi_{it}[x_t], \theta_{it}[x_t]/\phi_{it}[x_t])$ whose mean value is given by $\theta_{it}[x_t]$. We update this mean value recursively as:

$$\theta_{it}\left[x_t\right] = \alpha \theta_{i,t-1}\left[x_t\right] + (1-\alpha) E\left(\mathsf{v}_{i,t-1} \mid \mathbf{y}_{t-1}, x_t\right), \quad (11)$$

[1] If v is drawn from a gamma distribution whose shape parameter is very large, we can write $E(\log \mathsf{v}) \approx \log(E(\mathsf{v}))$.

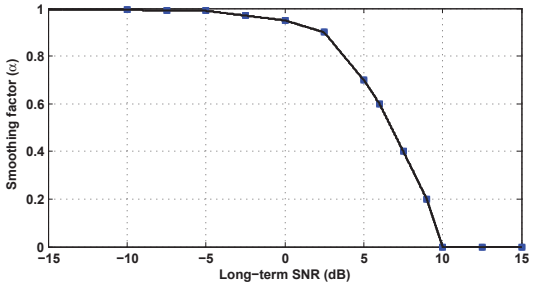

Fig. 3. An empirical α-SNR curve that is used in our experiments.

where the value of α controls the smoothing level to obtain the mean value for the prior distribution. Depending on the non-stationarity of the signals, different shape parameters are used for the speech and noise activations. Also, a single constant parameter is used for all the activations corresponding to one source. We learn this parameter at the end of the training stage by computing the average of all the shape parameters of the activation posterior distributions.

Our experiments show that the optimal amount of smoothing in (11) depends on the long-term input SNR. For low SNRs (high level of noise) a strong smoothing ($\alpha \rightarrow 1$) improves the performance by reducing unwanted fluctuations while for high SNRs a milder smoothing is preferred ($\alpha \rightarrow 0$). The latter case corresponds to obtaining the mean value θ directly using the information from the previous time frame. Here, in contrast to [4], we use a SNR-dependent value for the smoothing factor. Fig. 3 shows an $\alpha - \text{SNR}$ curve that was obtained in our computer simulations and is used in our experiments.

To calculate the long-term SNR from the noisy data, we implemented the approach from [15] that works well enough for our purpose. This approach assumes that the amplitude distribution of the clean speech signal is a gamma distribution with a small shape parameters (around 0.4) while the additive noise is assumed to be Gaussian. The noisy signal is assumed to be gamma-distributed whose shape parameter is uniquely determined from the long-term SNR. Promising results have been reported for different noise types [15].

3. EXPERIMENTS AND RESULTS

This section presents and discusses the results of our experiments. We evaluated the proposed approach for three noise types, hence, resulting to a BNMF-HMM with three states. We considered three non-stationary noises including factory and babble from NOISEX-92 database and city traffic from Sound Ideas database. The number of the basis vectors in the models were set using simulations on a small development set. We trained 100 basis vectors for each noise type. Also, we trained 60 basis vectors for a speaker-independent universal speech model using the training material from the TIMIT database.

We implemented a variety of NMF-based denoising algorithms for the purpose of comparison. We also considered a speech short-time spectral amplitude (STSA) estimator using super-Gaussian prior distributions in our experiments. These algorithms (three BNMF-based, two NMF-based, and one speech STSA estimator) are described in the following.

1. BNMF-HMM in which we used (8) where the underlying noise type in not known a priori
2. General-model BNMF that is a one-state BNMF-HMM where a single general noise dictionary with 200 basis vectors is learned using training data from all the considered noise types.
3. Oracle BNMF that is similar to BNMF-HMM but an oracle classifier is used to choose a noise model (and a pre-trained basis matrix) for enhancement instead of the proposed classifier. Therefore, this approach is an ideal case of the BNMF-HMM.
4. Oracle ML in which the multiplicative update rules of the KL-NMF in combination with a soft Wiener-type filtering is used to enhance the noisy signal. Similar to the oracle BNMF, the approach assumes that the noise type is known for the enhancement.
5. Oracle NHMM: this is basically the supervised causal NHMM in which the noise type is assumed to be given in advance. This approach is a modified version of [10] where we used a causal implementation of the method to denoise the signal. This modification is done to make the approach applicable for a scenario where we do not have access to future observations. To achieve causality, we simply replaced the forward-backward algorithm with forward algorithm in which the activations from the previous timestamps were used to initialize the current ones. We trained one universal NHMM (100 states with 10 basis each) for speech and one single-state NHMM for each noise type. The number of basis vectors for the noise models were set experimentally. We trained 100 basis vectors for the factory and city traffic noises and 30 basis vectors for the babble noise.
6. STSA-GenGamma [1], which is a speech STSA estimator using super-Gaussian priors. We used [5] to track the noise PSD, and we set $\gamma = \nu = 1$ since it is shown to be one of the best alternatives [1]. This algorithm is considered in our simulations as a state-of-the-art benchmark to compare NMF-based systems.

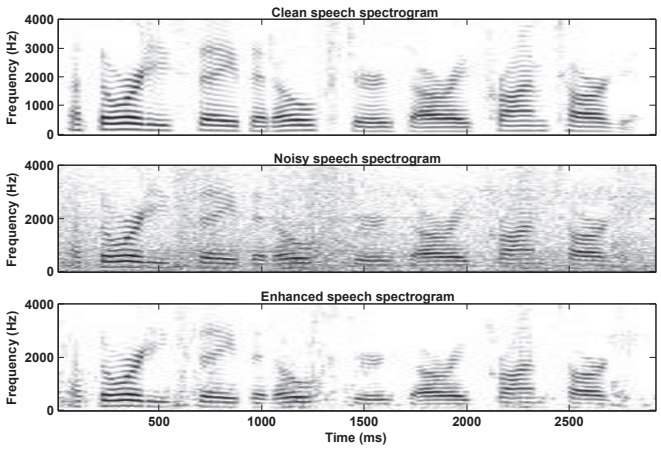

Fig. 4. Magnitude spectrogram of a sample clean speech, noisy speech, and enhanced speech signals in top, middle, and bottom panels, respectively. For this example, factory noise is added to the clean speech signal at 5 dB input SNR.

All the signals were down-sampled to 16-kHz in our experiments and the DFT was implemented using a frame length of 512 samples with 50% overlapped Hann windows. The signal synthesis was performed using the overlap-and-add procedure. The core test

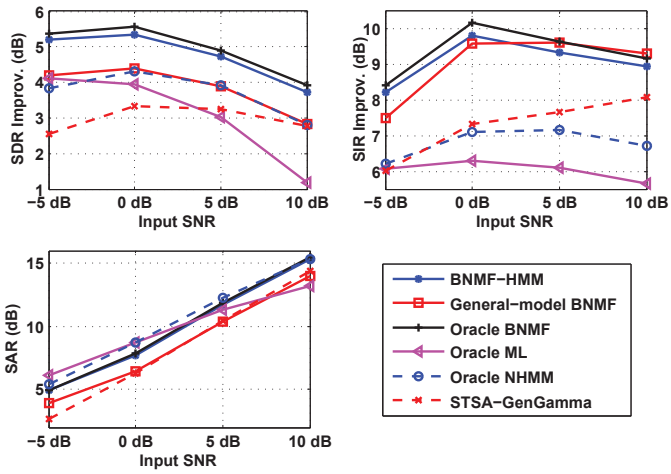

Fig. 5. SDR, SIR, and SAR values as objective measures to evaluate the denoising algorithms. We used "oracle" to point out that the noise type and its basis matrix is known for the speech enhancement. The evaluation is performed for 192 sentences and the results are averaged over different sentences and noise types. For the SDR and SIR, improvements gained by the enhancement systems are shown.

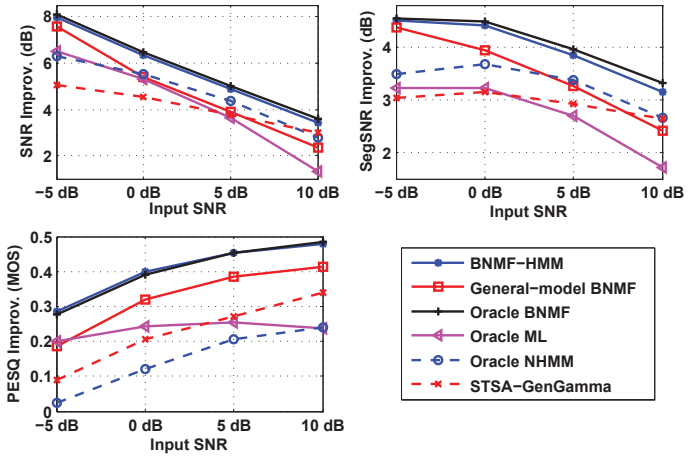

Fig. 6. long-term SNR, Segmental SNR (SegSNR), and PESQ improvements gained by the enhancement systems.

set of the TIMIT database (192 sentences) was used to evaluate the performance of the algorithms.

Fig. 4 demonstrates an example where BNMF approach is used to enhance a noisy signal. For this case, a female-uttered speech signal is degraded with factory noise at 5 dB input SNR. As the figure shows, the approach has reduced the noise significantly (a source to interference ratio (SIR) in the order of 11 dB) while the speech signal remains highly undistorted (a source to artifact ratio (SAR) of 13 dB).

Fig. 5 shows the source to distortion ratio (SDR), SIR, and SAR from BSS_Eval toolbox. A high value for all of these measures is desired. This figure shows SDR and SIR improvements obtained by the speech enhancement systems for clarity .

Our experiments show that BNMF-based algorithms are superior to the other methods. In particular, oracle BNMF and BNMF-HMM are the best algorithms. The difference in the performance of these two algorithms is marginal, which verifies that the proposed classification scheme is working successfully. We can also see that the NMF-based algorithms outperform the STSA-GenGamma in most of the cases. The only exception is that ML-NMF gives a worse SDR improvement at high input SNRs, which is mainly due to a small noise suppression (small SIR value). This is because the important temporal dependencies are ignored in ML-NMF. Moreover, the speech reconstruction rule, soft Wiener filtering, is not optimal in this case.

Another interesting result is that the oracle NHMM and general-model BNMF methods lead to similar SDR values. However, these two methods process the noisy signal differently. NHMM method does not suppress a lot of noise while it does not distort the speech signal either (i.e., SAR is high). This is reversed for the general-model BNMF.

Finally, Fig. 5 shows that general-model BNMF leads to a worse performance compared to the BNMF-HMM for which smaller noise-specific dictionaries are used. This result is in line with the other observations using supervised denoising algorithms [2]. This can be explained by noting that using a large noise dictionary increases the flexibility of the noise model, which in turn, increases the ambiguity in the nonnegative matrix factorization. It might be possible to get a better performance using a general noise model by adding more constraints into NMF, such as sparsity.

Fig. 6 shows our experimental results for long-term SNR, segmental SNR (SegSNR) that is limited to the range $[-10\text{dB}, 30\text{dB}]$, and perceptual evaluation of speech quality (PESQ). As it can be seen in the figure, the BNMF-based methods have led to the highest SNR, SegSNR and PESQ improvements. The figure shows that oracle BNMF and BNMF-HMM methods are superior to the other approaches and verify the results seen in Fig. 5.

Fig. 7 provides the result of another experiment where we studied the performance of the noise classification separately. To reduce the fluctuations and to have a clearer representation, (10) is smoothed over time and is depicted in this figure. For this experiment, speech was degraded by different noises (babble, factory, and city traffic) separately at 0 dB input SNR. Then, the classifier is applied to the noisy signal and the probability of each possible noise class (babble, factory, and city traffic) is computed and is shown in the figure. As it can be seen, the classifier works reasonably well in general. Most of the wrong classifications correspond to the case where the true noise is confused with babble noise. One reason for this confusion is perhaps 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. Though, as shown in Fig. 5 and Fig. 6, this confusion has reduced the noise reduction performance only marginally.

4. CONCLUSIONS

We presented an approach to integrate environmental noise classification and NMF-based speech enhancement. In the proposed method, acoustic environment is modeled using a discrete HMM where each state corresponds to one noise type. The derived MMSE estimate of the clean speech signal can be seen as a two step operator in which (1) the speech MMSE estimate is calculated for all the available noise types, and (2) the probability of each noise type is calculated, given the noisy observation, and is used to weight and sum the state-dependent MMSE estimates. Hence, the developed structure performs a simultaneous noise classification and speech enhancement and therefore does not require to know the noise type in

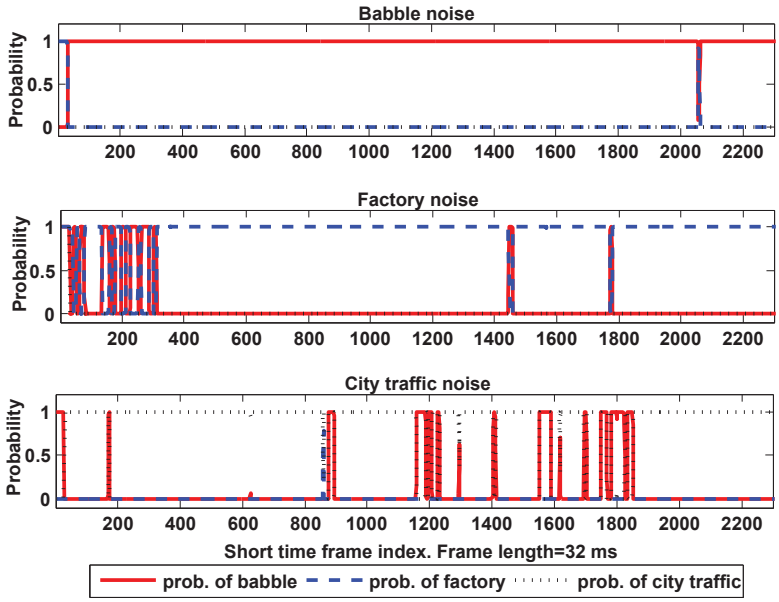

Fig. 7. Result of the noise classifier where (10) is smoothed over time and is plotted for a noisy signal at 0 dB input SNR. The underlying noise type is given in the titles of the subplots (which corresponds to babble, factory, and city traffic noises, respectively, from top to bottom). In each subplot, the probability of three noise classes (babble, factory, and city traffic noises) are shown.

advance. Our simulations show that the suggested system outperforms state-of-the-art and it is not restricted to know any a priori information that is difficult to obtain in practice.

References

[1] Jan S. Erkelens, Richard C. Hendriks, Richard Heusdens, and Jesper 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.

[2] Hossein Sameti, Hamid Sheikhzadeh, Li Deng, and Robert L. 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.

[3] Mikkel N. Schmidt and Jan 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.

[4] Nasser Mohammadiha, Jalil Taghia, and Arne 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.

[5] Richard C. Hendriks, Richard Heusdens, and Jesper 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.

[6] Kevin W Wilson, Bhiksha Raj, and Paris Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, 2008, pp. 411–414.

[7] Nasser Mohammadiha and Arne 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.

[8] L. Ma, D. J. Smith, and B. P. Milner, "Context awareness using environmental noise classification," in *European Conf. on Speech Communication and Technology (ISCA)*, 2003, pp. 2237–2240.

[9] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Process. Systems (NIPS)*. 2000, pp. 556–562, MIT Press.

[10] Gautham J. Mysore and Paris 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.

[11] Minje Kim and Paris Smaragdis, "Single channel source separation using smooth nonnegative matrix factorization with Markov random fields," in *IEEE Int. Workshop on Machine Learning for Signal Process. (MLSP)*, sep. 2013.

[12] Nasser Mohammadiha, Timo Gerkmann, and Arne 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 Acoust. (WASPAA)*, 2011, pp. 45–48.

[13] Ali Taylan Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009, Article ID 785152, 17 pages.

[14] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, feb. 1989.

[15] Chanwoo Kim and Richard 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.