SINGLE CHANNEL SPEECH ENHANCEMENT USING BAYESIAN NMF WITH RECURSIVE TEMPORAL UPDATES OF PRIOR DISTRIBUTIONS

Nasser Mohammadiha, Jalil Taghia, Arne Leijon *

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

ABSTRACT

We present a speech enhancement algorithm which is based on a Bayesian Nonnegative Matrix Factorization (NMF). Both Minimum Mean Square Error (MMSE) and Maximum a-Posteriori (MAP) estimates of the magnitude of the clean speech DFT coefficients are derived. To exploit the temporal continuity of the speech and noise signals, a proper prior distribution is introduced by widening the posterior distribution of the NMF coefficients at the previous time frames. To do so, a recursive temporal update scheme is proposed to obtain the mean value of the prior distribution; also, the uncertainty of the prior information is governed by the shape parameter of the distribution which is learnt automatically based on the nonstationarity of the signals. Simulations show a considerable improvement compared to the maximum likelihood NMF based speech enhancement algorithm for different input SNRs.

Index Terms— Speech enhancement, NMF, MMSE, MAP

1. INTRODUCTION

Single channel speech enhancement has been a research topic for a long time. A major outcome of these techniques is the improved quality and reduced listening effort in the presence of a strong interfering noise signal. In this paper, we study a Nonnegative Matrix Factorization (NMF) based speech enhancement approach.

NMF is a popular factorization method which projects the given nonnegative matrix onto its nonnegative basis vectors. NMF has been used in a variety of applications in audio processing including, but not limited to, blind source separation [1, 2, 3] and speech enhancement [4, 5, 6]. For these applications, the magnitude (or power) spectrogram of the speech signal, X, is factorized into its basis matrix T and a time-varying NMF coefficients matrix V, such that: $X \approx TV$.

The temporal dependencies of the audio signals is ignored in the basic NMF. Moreover, an important challenge in NMF based speech enhancement approaches is that the basis matrices for speech and noise may be quite similar. Using the temporal continuity of the underlying sources has been shown to be promising to overcome this problem. In [4], it is assumed that log of the NMF coefficients corresponding to each basis vector follow a Normal distribution; hence, a regularized NMF is proposed to enhance the noisy speech which has separate training and enhancement stages. The statistics of the Normal distributions are assumed to be fixed and are obtained in the end of the training stage. Another approach is proposed in [5] in which the temporal activity of the basis vectors are modeled using a hidden Markov model; in this approach, speech is modeled using multiple nonnegative dictionaries, and noise is modeled using a single dictionary.

In our previous work [6], a linear MMSE filter was proposed to enhance the noisy speech; though, no explicit prior density function was assumed for the parameters of interest. In this paper, we explicitly assign a prior distribution, in the form of a Gamma density function, to the NMF coefficients. From the available Bayesian NMF approaches, e.g. [7, 8], we use [7] and extend the proposed Bayesian NMF to model the noisy signal, and we derive the MMSE filter to estimate the clean speech component. To employ the time continuity of the speech and noise signals, the posterior distribution of the NMF coefficients at the previous time frames are widened and are used as the prior distribution for the current time frame. Finally, we derive the Maximum a-Posteriori (MAP) estimator using the aforementioned prior distribution. The proposed estimators lead to superior results for the evaluated noise types compared to the competing algorithms. The rest of the paper is organized as follows: In Section 2, the exploited probabilistic NMF is reviewed. The speech enhancement framework is outlined in Section 3. In Sections 4 and 5, the MMSE and MAP estimators are derived; performance evaluation is given in Section 6.

2. REVIEW OF PROBABILISTIC NMF

In this section, we review the probabilistic NMF which is proposed in [7]. The following model is considered in the aforementioned to perform NMF:

$$\begin{aligned} X(k,\tau) &= \textstyle\sum_i Z(k,i,\tau), \\ Z(k,i,\tau) &\sim \mathcal{PO}\left(Z(k,i,\tau);\ \Lambda(k,i,\tau)\right), \end{aligned} \tag{1}$$

where $Z(k,i,\tau)$ are latent sources, $\mathcal{PO}(z;\lambda) = \exp(z\log\lambda - \lambda - \log\Gamma(z+1))$, with $\Gamma(z+1) = z!$ as the Gamma function, denotes the Poisson distribution, and $\Lambda(k,i,\tau) = T(k,i)\,V(i,\tau)$. Consequently, $X(k,\tau)$ is assumed to have a Poisson distribution with mean value equal to $\sum_i \Lambda(k,i,\tau)$. For the speech enhancement purposes, $X(k,\tau) = |Y(k,\tau)|$ where $Y(k,\tau)$ denotes the DFT coefficient for frequency bin k and time-frame τ of the noisy signal.

2.1. NMF using Maximum Likelihood (ML) Estimation

An Expectation Maximization (EM) algorithm is proposed in [7] to find the ML estimate of T and V in (1). In the expectation step, the expected values of the latent sources, conditioned on X, T and V ($\bar{Z}(k,i,\tau)$) are calculated as:

$$\bar{Z}(k,i,\tau) = X(k,\tau)\,\frac{\Lambda(k,i,\tau)}{\sum_{i'}\Lambda(k,i',\tau)}. \tag{2}$$

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

The maximization step involves maximizing of

$$Q = \sum_{k,i,\tau} \left(-\Lambda(k,i,\tau) + \bar{Z}(k,i,\tau) \log\left(\Lambda(k,i,\tau)\right) \right), \quad (3)$$

w.r.t. T and V, which after combining with (2) results to the following multiplicative rules for T and V that we refer to as ML-NMF:

$$\begin{aligned} T(k,i) &\leftarrow T(k,i) \frac{\sum_{\tau} V(i,\tau) \left(X(k,\tau) / \sum_{i'} \Lambda(k,i',\tau) \right)}{\sum_{p} V(i,p)}, \\ V(i,\tau) &\leftarrow V(i,\tau) \frac{\sum_{k} T(k,i) \left(X(k,\tau) / \sum_{i'} \Lambda(k,i',\tau) \right)}{\sum_{q} T(q,i)}, \end{aligned} \quad (4)$$

these updates are performed in an iterative manner until convergence.

2.2. Variational Bayes for Bayesian Inference

In the Bayesian framework, T and V are considered as some random variables; the following Gamma prior distributions are considered for the basis elements $T(k,i)$ and NMF coefficients $V(i,\tau)$:

$$\begin{aligned} V(i,\tau) &\sim \mathcal{G}\left(V(i,\tau); a(i,\tau), b(i,\tau)/a(i,\tau)\right), \\ T(k,i) &\sim \mathcal{G}\left(T(k,i); c(k,i), d(k,i)/c(k,i)\right), \end{aligned} \quad (5)$$

in which $\mathcal{G}(x;k,\theta) = \exp((k-1)\log x - x/\theta - \log\Gamma(k) - k\log\theta)$ denotes the Gamma density function with θ as the scale parameter. Since the exact Bayesian inference turns out to be difficult, a Variational Bayes has been proposed in [7]. Hence, in an iterative scheme the current parameters of the posterior distributions of $Z(k,:,\tau)$ are used to update the parameters of the posterior distributions of $T(k,i)$ and $V(i,\tau)$, and these new parameters are used to update the posterior distributions of $Z(k,:,\tau)$ in the next iteration (':' denotes 'all the indices'). The iterations are carried on until convergence. The posterior distributions for $Z(k,:,\tau)$ are multinomial density functions, while for $T(k,i)$ and $V(i,\tau)$ they are Gamma density functions. Here, we only mention the update equations of the expected values of $Z(k,:,\tau)$ as we need them later:

$$E\left(Z(k,i,\tau) \mid X\right) = \frac{e^{E(\log(T(k,i)V(i,\tau))|X)}}{\sum_{i=1} e^{E(\log(T(k,i)V(i,\tau))|X)}} X(k,\tau). \quad (6)$$

Full details of this framework can be found in [7].

3. SPEECH ENHANCEMENT FRAMEWORK

In the rest of the document, we denote the τth column of a matrix X by $\mathbf{x}_{.\tau}$. Our algorithm consists of training and enhancement steps. For both steps, the given time domain signal is segmented, windowed, and transformed into the frequency domain to obtain the spectrogram. During the training step, where the training clean speech and noise signals are given, NMF is applied to the magnitude spectrogram of the clean speech signal $|S|$ (where $|\cdot|$ is used to show the element-wise absolute value) and noise signal $|N|$ to obtain the speech basis matrix, T_S, and noise basis matrix, T_N. For the enhancement, an overlap-add framework is utilized to process each frame of the noisy speech, $\mathbf{y}_{.\tau}$, separately. The enhancement step depends on the estimator type which is discussed in the following sections.

4. MMSE ESTIMATOR

We model the magnitude spectrogram of the clean speech and noise signals by (1). Using the Bayesian framework outlined in Section 2.2, the posterior distributions of the speech and noise basis matrices, T_S and T_N, are found using the training data. For the enhancement, a complete basis matrix is built as $T = [T_S\ T_N]$ with the concatenated distribution parameters of T_S and T_N. For simplicity we assume that the training sets are so rich that the distribution of T remains constant for the period of the test segment. We assume that the magnitude spectrogram of the noisy speech signal is approximated by the sum of the speech and noise magnitude spectrograms, i.e., $|\mathbf{y}_{.\tau}| \approx |\mathbf{s}_{.\tau}| + |\mathbf{n}_{.\tau}|$. Note that although this assumption is not theoretically well justified, it is a common assumption in NMF based speech processing [1, 2, 4, 5, 6], and has led to good results in practice. Let us denote the number of the basis vectors for speech as I and for noise as J; the first line of (1) is now written as: $|Y(k,\tau)| = \sum_{i=1}^{I+J} Z(k,i,\tau)$. Considering a proper prior distribution for the NMF coefficients (see Section 4.1), the Bayesian framework is applied to $|\mathbf{y}_{.\tau}|$ to find the posterior distributions of the NMF coefficients $U(i,\tau)$ and latent sources $Z(k,i,\tau)$.

The MMSE estimate [9, Sec 11.4] of the magnitude of the speech DFT coefficients is given by the conditional expectation of the sum of the corresponding latent sources as:

$$\widehat{|S(k,\tau)|} = E\left(|S(k,\tau)| \mid |\mathbf{y}_{.\tau}|\right) = E\left(\sum_{i=1}^{I} Z(k,i,\tau) \mid |\mathbf{y}_{.\tau}|\right). \quad (7)$$

Exploiting the properties of the conditional expectation, the MMSE estimate in (7) is written as:

$$\widehat{|S(k,\tau)|} = \sum_{i=1}^{I} E\left(Z(k,i,\tau) \mid |\mathbf{y}_{.\tau}|\right). \quad (8)$$

From (6), we also have:

$$E\left(Z(k,i,\tau) \mid |\mathbf{y}_{.\tau}|\right) = \frac{e^{E(\log(T(k,i)U(i,\tau))||\mathbf{y}_{.\tau}|)}}{\sum_{i=1}^{I+J} e^{E(\log(T(k,i)U(i,\tau))||\mathbf{y}_{.\tau}|)}} |Y(k,\tau)|. \quad (9)$$

Inserting (9) in (8), we obtain:

$$\widehat{|S(k,\tau)|} = \frac{\sum_{i=1}^{I} e^{E(\log(T(k,i)U(i,\tau))||\mathbf{y}_{.\tau}|)}}{\sum_{i=1}^{I+J} e^{E(\log(T(k,i)U(i,\tau))||\mathbf{y}_{.\tau}|)}} |Y(k,\tau)|. \quad (10)$$

The time domain enhanced signal is reconstructed using the noisy phase information.

4.1. Assigning Informative Priors for Bayesian NMF

During the training, we assign some sparse and broad prior distributions to T and V according to (5). For this purpose, c and d are chosen such that the mean of the prior distribution for T is small and its variance is very high. On the other hand, a and b are chosen such that the prior distribution of V has a mean corresponding to the scale of the data and a high variance to represent uncertainty. To have good initializations for the Variational Bayes approach, the ML-NMF from (4) is applied first (e.g. for 10 iterations), and the obtained estimates of T and V are used as the initial mean values for the posterior distributions of T and V.

In the enhancement stage, to use the temporal correlation of the noise and speech signals, we use a data-driven prior for the NMF

coefficients U. To do so and also to account for the nonstationarity of the signals, a proper prior for U is obtained by widening the posterior distributions of U from the previous time frames. Hence, the posterior distributions of U at $t = 0, 1, \ldots, \tau - 1$ are used to obtain a prior distribution for time frame $t = \tau$. Denoting this prior distribution as $U(i,\tau) \sim \mathcal{G}(U(i,\tau); a(i), b(i,\tau)/a(i))$, we have:

$$E(U(i,\tau)) = b(i,\tau), \quad \frac{\sqrt{var(U(i,\tau))}}{E(U(i,\tau))} = \frac{1}{\sqrt{a(i)}}. \quad (11)$$

We assign the following recursively updated mean value, which is conditioned on all the observations at $t = 0, 1, \ldots, \tau - 1$, for the prior distribution:

$$b(i,\tau) = \alpha E\left(U(i,\tau-1) \mid \left|\mathbf{y}_{.(\tau-1)}\right|\right) + (1-\alpha) b(i,\tau-1), \quad (12)$$

here, the value of α controls the smoothing level to obtain the prior. In (11), different shape parameters are used for the speech and noise NMF coefficients, i.e., $a(1:I) = a_{speech}$ and $a(I+1:I+J) = a_{noise}$. In this form of prior, the ratio between the standard deviation and the expected value is the same for all the NMF coefficients of a source, independent of time. The shape parameter a represents the uncertainty of the prior which in turn corresponds to the stationarity of the signal being processed. We can learn this parameter in the training stage using the clean speech or noise signals. For this purpose, at the end of the training stage, the shape parameter of the posterior distributions of all the NMF coefficients are calculated and their mean value is taken for this purpose.

5. MAP ESTIMATOR

In this section, we derive the MAP estimate of the magnitude of the speech DFT coefficients. For this approach, the maximum likelihood estimates of the speech and noise basis matrices, T_S and T_N, are obtained during the training step using (4). Now, the basis matrix for the observed noisy speech, T, is obtained as $T = (T_S\ T_N)$, and it is kept fixed during the enhancement. In the enhancement step, to perform NMF as $|\mathbf{y}_{.\tau}| \approx T\mathbf{u}_{.\tau}$, we assign a Gamma prior distribution as $\mathcal{G}(U_p(i,\tau); a(i), b(i,\tau)/a(i))$ to the NMF coefficient $U_p(i,\tau)$. In the following, we aim to find the MAP estimate of the NMF coefficients, denoted by $U(i,\tau)$, using the mentioned prior distribution. We use the EM algorithm for this purpose. The expectation step is identical to Section 2.1 [10, p.454], so $\bar{Z}(k,i,\tau) = |Y(k,\tau)|\,\Lambda(k,i,\tau)/\sum_{i'}\Lambda(k,i',\tau)$ where $\Lambda(k,i,\tau) = T(k,i)\,U(i,\tau)$. In the maximization step, we maximize an objective function Q which is the sum of two terms: the first term corresponds to the likelihood of the observations (3), and the second term involves the likelihood of the NMF coefficients under the given prior distribution:

$$\begin{aligned} Q = &\sum_{k,i,\tau}\left(-\Lambda(k,i,\tau) + \bar{Z}(k,i,\tau)\log\left(\Lambda(k,i,\tau)\right)\right) \\ &+ \sum_{i,\tau}\left((a(i)-1)\log U(i,\tau) - \tfrac{U(i,\tau)}{b(i,\tau)}\right), \end{aligned} \quad (13)$$

Taking derivative of (13) w.r.t. $U(i,\tau)$ and setting it to zero yields the following iterative update rule for $U(i,\tau)$:

$$\begin{aligned} L(i,\tau) &= U(i,\tau)\sum_k T(k,i)\left(|Y(k,\tau)|\,/\sum_{i'}\Lambda(k,i',\tau)\right), \\ U(i,\tau) &\leftarrow \frac{\max\left(L(i,\tau)+a(i)-1,\epsilon\right)}{\sum_q T(q,i)+1/b(i,\tau)}, \end{aligned} \quad (14)$$

where ϵ is a small positive number to ensure that we get nonnegative values for U. It is straightforward to show that this stationary point is the maximum if $a(i) \geq 1$. (14) is performed iteratively until convergence. The enhanced speech component can be obtained now using a Wiener type gain as:

$$\widehat{|S(k,\tau)|} = \frac{\sum_{i=1}^{I}\Lambda(k,i,\tau)}{\sum_{i=1}^{I+J}\Lambda(k,i,\tau)}|Y(k,\tau)|. \quad (15)$$

Choosing $a(i) = 1$ and $b(i,\tau) = \infty$, (14) results to the ML estimate of the NMF coefficient $U(i,\tau)$. In this case, applying (15) will result to the speech estimate using the ML-NMF which is considered in the simulations.

5.1. Assigning Informative Priors for MAP Estimator

We assign a prior distribution to NMF coefficients in the form of $U_p(i,\tau) \sim \mathcal{G}(U_p(i,\tau); a(i), b(i,\tau)/a(i))$, for which

$$E(U_p(i,\tau)) = b(i,\tau), \quad \frac{\sqrt{var(U_p(i,\tau))}}{E(U_p(i,\tau))} = \frac{1}{\sqrt{a(i)}}. \quad (16)$$

The mean value $b(i,\tau)$ is obtained by recursively smoothing the MAP estimates of the NMF coefficients at $t = 0, 1, \ldots, \tau - 1$ as:

$$b(i,\tau) = \beta U(i,\tau-1) + (1-\beta)\, b(i,\tau-1). \quad (17)$$

To have an estimate of the shape parameter a, we may use the NMF coefficients matrix V at the end of the training stage. Estimating the mean and variance of all the elements of V and calculating their ratio as the shape parameter did not result to a reasonable number; the reason might be that in each time frame only some of the coefficients have high contribution to explain the data and only their variation is important. Hence, we project the signal onto its basis vectors and measure the variation of the effective coefficients. Therefore, at the end of the training step, each time frame of the training noise signal, $|\mathbf{n}_{.\tau}|$, is projected onto each of the noise basis vectors by simple inner product normalized by the norm of the basis vector:

$$P(i,\tau) = \frac{|\mathbf{n}_{.\tau}|^{\top}\mathbf{t}_{.i}}{\mathbf{t}_{.i}^{\top}\mathbf{t}_{.i}},$$

where $\mathbf{t}_{.i}$ is the ith basis vector, ith column of the basis matrix T_N. Then, m highest values are chosen (we use $m = 5$), and their average value is calculated and smoothed over the surrounding M frames (a simple averaging is used for this purpose) which is represented by $\mu_P(t)$. M is the number of the frames in which the signal is assumed to be stationary. Finally, the variance of the m selected $P(i,\tau)$ is calculated in the entire block of M frames and is called $\sigma_P^2(t)$. The parameter a_{noise} is now obtained as:

$$\mathrm{a}(\mathrm{t}) = \frac{\mu_P^2(t)}{\sigma_P^2(t)}, \qquad a_{noise} = \frac{1}{N}\sum_{t=1}^{N}\mathrm{a}(\mathrm{t}), \quad (18)$$

where N is the number of the frames of the training data. a_{speech} is obtained using a similar procedure. In practice, the result of (18) was similar to the estimate obtained in Section 4.1, which in general may need some tuning to get the best results for the enhancement.

6. EVALUATIONS

We evaluate and compare the proposed MMSE and MAP estimators with the ML-NMF, LMMSE estimator from [6], and a Wiener filtering in which the noise PSD was estimated using [11]. The implementation of the LMMSE and Wiener filters, and data preparation were done similarly to [6]. We used speech from the Grid Corpus [12] and noise from the NOISEX-92 databases. All the signals were

Table 1: SDR improvements in dB averaged over the babble and factory noises

Input *SNR*	ML-NMF	LMMSE	Wiener	MAP	MMSE
0 dB	1.2	**2.7**	1.5	2.4	2.4
5 dB	0.7	2.7	1.7	3.1	**3.2**
10 dB	-1	1.7	1.1	1.8	**2.5**

Table 2: Segmental speech *SNR* (SNR_{sp}) and Segmental Noise Reduction (*SegNR*) in dB averaged over the babble and factory noises

		ML-NMF	LMMSE	Wiener	MAP	MMSE
0 dB	SNR_{sp}	6.8	8.6	4.9	8.9	10.1
	SegNR	6.8	6.7	13.2	6.3	5.6
5 dB	SNR_{sp}	7.8	10	8.6	9.5	11.1
	SegNR	6.6	6.7	10.4	8.5	7.2
10 dB	SNR_{sp}	8.9	11.4	12.7	10.2	12.1
	SegNR	6.3	6.7	7.9	10.4	8.6

down-sampled to 16 kHz. The speech was degraded by adding babble or factory noise at three different *SNR*s: 0 dB, 5 dB, and 10 dB. The data was scaled properly to reduce the rounding error of the Bayesian framework which is a side effect of the Poisson model assumption. The overlap-add approach with Hann window was utilized to process the noisy speech; the time frames had a length of 512 samples with 50% overlap. A separate model was trained for each noise type, and one speaker independent model was trained for the speech signal; separate train and test sets were used for both speech and noise signals. 80 sentences (10 sentences from 4 male and 4 female speakers) were considered in the test set. The results are averaged over all the test set and both noise types. 60 basis vectors for speech and 100 basis vectors for each noise were trained. $a_{factory} = 10$ and $a_{babble} = 8$ were obtained and used as the shape parameters for both MMSE and MAP estimators for the factory and babble noises, respectively. Even though using the approach from Section (4.1), or (5.1), we get $a_{speech} \approx 5$, applying the corresponding prior distribution did not improve the performance; hence, in our simulations we set $a_{speech} = 1$ and $b_{speech} = \infty$ to use an uninformative prior for speech in (13).

The performance of the speech enhancement algorithms are evaluated using the Source to Distortion Ratio (*SDR*); From the three measures introduced in [13], we present only *SDR* due to lack of space; *SDR* represents the overall quality of the enhanced speech where "having no artifacts" and "noise reduction" are equally important. We also compare the algorithms using segmental speech*SNR* (SNR_{sp}), and segmental noise reduction (*SegNR*) which are calculated in a shadow filtering framework as [6, 14]. These measures are presented to study the "introducing artifacts" and "reducing noise" aspects of the enhancement algorithms, separately.

Table 1 shows the *SDR* improvements in dB which is averaged over the babble and factory noises. MMSE estimator results to the highest improvements for high input *SNR*s, while results of the LMMSE approach are better for 0 dB input *SNR*. MMSE estimator results to slightly better *SDR* improvements than the MAP estimator. Both MMSE and MAP filters give superior results compared to the ML-NMF and Wiener filtering. We found the smoothing coefficient α (Eq. (12)) effective for the enhancement performance. Choosing $\alpha = 1$, which means having no smoothing to obtain the mean value for the prior distribution, results to better *SDR* improvements for high input *SNR*s but degrades the performance for low input *SNR*s. In contrast, increasing the smoothing level will improve the results for low input *SNR*s but degrade it for high input *SNR*s. We used $\alpha = 0.6$ which gave a compromise for both low and high input *SNR*s. This behavior was not observed for the MAP estimator, and we used $\beta = 0.9$ (Eq. (17)). Table 2 shows the segmental speech*SNR* (SNR_{sp}), and segmental noise reduction (*SegNR*), averaged over both the babble and factory noises, at different input *SNR*s. For both measures a high value is desired, and SNR_{sp} is inversely proportional to the speech distortion. MMSE estimator results to higher SNR_{sp} than MAP estimator, but this is reversed for *SegNR*. Compared to LMMSE, mostly MMSE gives higher SNR_{sp} and *SegNR*. These results were confirmed by informal listening. Some audio examples are available at `http://www.ee.kth.se/~nmoh/se_using_bnmf`.

In terms of the complexity, MMSE estimator requires more complicated training procedure than the ML and MAP estimators. However, the computational time for the enhancement step is quite similar for all of them.

7. CONCLUSION

We presented a speech enhancement algorithm using a Bayesian NMF framework. We derived both the MMSE and MAP estimators. To exploit the time continuity of the noise signals, we used data-driven prior distributions for the noise NMF coefficients. The proposed estimators were evaluated using the Source to Distortion Ratio (*SDR*). The MMSE estimator was found to be the superior which outperformed the standard NMF algorithm by 2.4 dB higher *SDR* improvement, averaged over all the evaluated input *SNR*s and noise types. The MAP estimator led to similar results as the MMSE at low input *SNR*s, but its performance was worse for higher input *SNR*s.

8. REFERENCES

[1] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. ASLP*, vol. 15, pp. 1–12, 2007.

[2] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. ASLP*, vol. 15, no. 3, pp. 1066–1074, 2007.

[3] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. ASLP*, vol. 18, no. 3, pp. 550–563, March 2010.

[4] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Interspeech*, 2008, pp. 411–414.

[5] 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 *IEEE Int. Conf. ICASSP*, 2011.

[6] 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.

[7] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," Cambridge University Engineering Department, Tech. Rep. CUED/F-INFENG/TR.609, 2008.

[8] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *International Conference on Machine Learning*, 2010.

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.

[10] C. M. Bishop, *Pattern Recognition and machine learning*. Springer, 2006.

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

[12] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *JASA*, vol. 120, pp. 2421–2424, 2006.

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[14] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, pp. 1110–1126, 2005.