TRANSIENT NOISE REDUCTION USING NONNEGATIVE MATRIX FACTORIZATION

Nasser Mohammadiha, Simon Doclo

Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all
University of Oldenburg, Germany

ABSTRACT

Reducing highly non-stationary transient noise, such as keyboard typing noise, remains a challenging problem for many single-channel speech enhancement algorithms. This paper proposes two approaches based on nonnegative matrix factorization (NMF) and probabilistic latent component analysis for transient noise reduction using a pre-trained transient noise dictionary and a universal speaker-independent speech dictionary. In addition, we develop an NMF-based speech enhancement scheme to simultaneously reduce transient and non-transient background noise, in which a low-dimensional dictionary is learnt from the noisy observations to model the background noise. We exploit the temporal dependencies of speech and background noise to design and apply informative priors via a probabilistic framework, while ignoring the temporal dynamics of the transient noise. Experimental results show that the proposed algorithms can improve the perceptual evaluation of speech quality (PESQ) up to 1.2 MOS for the keyboard typing noise.

Index Terms— nonnegative matrix factorization, probabilistic latent component analysis, impulsive noise, transient noise

1. INTRODUCTION

Single-channel speech enhancement is a widely investigated problem, which however has not been completely solved. Many speech enhancement approaches have been developed in the past that work quite well for stationary noise, e.g., spectral enhancement approaches [1, 2]. These methods typically require an estimate of the noise power spectral density (PSD), which is difficult to obtain if the noise is highly non-stationary [3].

Reducing transient noise, such as the sound of keyboard typing or machinegun noise, is a noise reduction problem where the noise characteristics vary rapidly, and therefore, accurately estimating the noise PSD is very difficult. Accordingly, traditional spectral enhancement schemes can not be employed to effectively reduce transient noise and alternative approaches have been developed for this purpose. For example, a transient noise pulse removal system has been proposed in [4], where the presence of the noise is first detected and then a matched filter is used to remove the noise. Another phase-based tapping noise detection and suppression approach has been proposed in [5].

Alternatively, supervised noise reduction methods can provide a promising framework to reduce transient interference. In these methods, the characteristics of different noise types are first learnt using training samples and are then used to denoise an observed noisy signal. Following this idea, a supervised method has been proposed in [6], where the noise spectral vectors from some training data are used to construct an affinity matrix between the noise and the noisy observation. A low-rank approximation of this matrix is then used to obtain an estimate of the noise spectral component. In this paper, we develop supervised transient noise reduction methods using nonnegative matrix factorization (NMF) [7], where the speech, the transient noise, and the remaining background noise signals are modeled by low-rank representations.

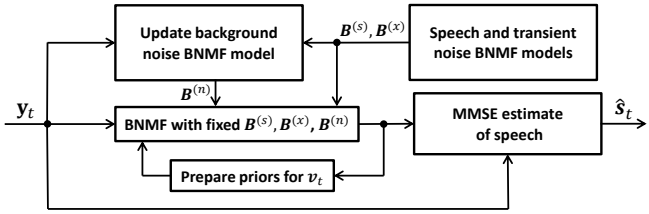


Fig. 1: Block diagram of transient and background noise reduction using Bayesian NMF. The noisy magnitude spectral vector $\mathbf{y}_t$ is approximated as $\mathbf{B}\mathbf{v}_t$, where $\mathbf{B}$ and $\mathbf{v}_t$ denote the NMF dictionary and NMF coefficients vector, respectively.

Recently, NMF-based approaches using probabilistic latent component analysis (PLCA) [8] and a Bayesian formulation of NMF (BNMF) [9] have been developed to exploit temporal dependencies of the signals. In [9], the posterior distributions of the dictionaries are obtained using speech and noise training data. The clean speech DFT coefficients are then obtained by combining a minimum mean square error (MMSE) estimate of the speech DFT magnitudes with the noisy phase. An important aspect of this approach is that temporal dependencies are used to construct informative priors for the NMF coefficients (the coefficients corresponding to the NMF dictionary elements). Experimental results in [9] showed that for noise types that are more stationary than the speech signal, e.g., babble or traffic noise, a flat prior for the speech NMF coefficients gives a better performance. Hence, informative priors were only used for the noise NMF coefficients [9].

In this paper, we propose supervised PLCA and BNMF based denoising approaches to reduce highly non-stationary transient noise. In addition, we develop a new method to simultaneously reduce transient and non-transient background noises using NMF. Fig. 1 shows a block diagram of the proposed BNMF-based speech enhancement approach, where the transient noise and (speaker-independent) speech NMF models are learnt offline, while the background noise NMF model is learnt online from the noisy observations. Both proposed methods exploit the temporal dynamics of the speech and noise signals, where we use the realistic assumption that transient noise is more non-stationary than speech, which in turn is assumed to be more non-stationary than the background noise. Accordingly, we use flat priors for the transient noise NMF

This research was supported by the Cluster of Excellence 1077 "Hearing4all", funded by the German Research Foundation (DFG).

coefficients, while for the background noise and the speech (in contrast to [9]) we exploit temporal correlations to set informative priors. This process is in principle similar to the bandpass filtering in RASTA [10] and the approach used in [11]. Our experiments with keyboard and machinegun transient noises at different signal-to-noise ratios show that both proposed methods lead to a good speech enhancement performance, where the BNMF scheme outperforms the PLCA approach.

2. SIGNAL MODEL

We consider a single-channel noise reduction problem with additive highly non-stationary transient and relatively more stationary background noise. All signals are transformed to the time-frequency domain by applying the discrete Fourier transform (DFT) to short (overlapping) signal frames. Let y_{kt} denote the noisy DFT magnitude at frequency bin k and time frame t. Similarly, let s_{kt}, n_{kt}, and x_{kt} denote the clean speech, background noise, and transient noise DFT magnitudes, respectively. Similarly to [8, 9], we assume that the DFT magnitudes of speech and noise add up to obtain the noisy DFT magnitudes. This can be formulated in a vector form as:

$$\mathbf{y}_t = \mathbf{s}_t + \mathbf{n}_t + \mathbf{x}_t, \tag{1}$$

where $\mathbf{y}_t = [y_{1t}, y_{2t} \dots y_{Kt}]^\top$, $\top$ denotes transpose, and K is the number of frequency bins. $\mathbf{s}_t$, $\mathbf{n}_t$, and $\mathbf{x}_t$ are defined similarly.

We use NMF to approximate each of these signals by a low-rank representation. For example, $\mathbf{S} \approx \mathbf{B}^{(s)}\mathbf{V}^{(s)}$, where $\mathbf{S}$ is the $K \times T$ speech magnitude spectrogram where T denotes the number of speech training frames, $\mathbf{B}^{(s)}$ is the $K \times I^{(s)}$ speech dictionary, and $\mathbf{V}^{(s)}$ is the $I^{(s)} \times T$ speech NMF coefficient matrix. The model order $I^{(s)}$ is chosen such that $I^{(s)} < \min(K, T)$ and thus the factorization yields a low-rank approximation of the input matrix. The noisy DFT magnitudes can now be approximated as

$$\mathbf{y}_t \approx \left[\mathbf{B}^{(s)}\ \mathbf{B}^{(n)}\ \mathbf{B}^{(x)}\right]\left[\mathbf{v}_t^{(s),\top}\ \mathbf{v}_t^{(n),\top}\ \mathbf{v}_t^{(x),\top}\right]^\top = \mathbf{B}\mathbf{v}_t, \tag{2}$$

where $\mathbf{B}^{(n)}$ and $\mathbf{B}^{(x)}$ represent the background noise and the transient noise dictionaries, and $\mathbf{v}_t$ denotes the NMF coefficient vector corresponding to the t-th frame. All vectors $\mathbf{v}_t$ for $t = 1, \dots T$ are stacked in a matrix denoted by $\mathbf{V}$.

In this paper, we are interested to obtain an estimate of the clean speech DFT magnitude $\mathbf{s}_t$ given $\mathbf{y}_t$, and the speech and transient noise dictionaries $\mathbf{B}^{(s)}$, and $\mathbf{B}^{(x)}$. We first learn these dictionaries using speech and transient noise training data, whereas the NMF coefficient vector $\mathbf{v}_t$ and the background noise dictionary $\mathbf{B}^{(n)}$ are updated online. After computing an estimate of the clean speech magnitude $\mathbf{s}_t$, the enhanced speech signal is obtained using the noisy phase, inverse DFT, and overlap-and-add framework.

3. NMF BASED SPEECH ENHANCEMENT

The proposed speech enhancement algorithms using PLCA and BNMF are explained in this section. In subsections 3.1 and 3.2, we assume that the dictionary $\mathbf{B}$ is given, while in subsection 3.3, we explain how $\mathbf{B}$ is obtained and updated over time.

3.1. Enhancement Strategy using PLCA

We use the dynamic PLCA approach proposed in [8], which is based on the following state-space representation:

$$E(\mathbf{v}_t) = \mathbf{A}\hat{\mathbf{v}}_{t-1}, \tag{3}$$

$$\mathbf{y}_t \sim \text{multinomial}(\mathbf{B}\mathbf{v}_t), \tag{4}$$

where $\mathbf{A}$ is the autoregressive coefficient matrix that is learnt similarly to [8], $E(\cdot)$ denotes the expected value, $\hat{\mathbf{v}}_{t-1}$ is the estimated NMF coefficient vector at time $t-1$, and $\sim$ indicates that $\mathbf{y}_t$ is sampled from the multinomial distribution with the given parameter.

Following [8], we use a filtering approach to estimate $\mathbf{v}_t$ that is written as:

$$\hat{\mathbf{v}}_t = \frac{(\mathbf{A}\hat{\mathbf{v}}_{t-1})^{\boldsymbol{\beta}} \odot \tilde{\mathbf{v}}_t}{\sum_i (\mathbf{A}\hat{\mathbf{v}}_{t-1})^{\boldsymbol{\beta}} \odot \tilde{\mathbf{v}}_t}, \tag{5}$$

where $(\cdot)^{\boldsymbol{\beta}}$ and $\odot$ denote element-wise power and product operators, respectively, $\tilde{\mathbf{v}}_t$ is the obtained NMF coefficient vector by applying standard PLCA [12] on $\mathbf{y}_t$, and $\boldsymbol{\beta}$ specifies the signal-dependent prior strength. We use a single parameter $\beta^{(s)} < 1$ for all speech NMF coefficients while we ignore the temporal dynamics of the transient noise signal and set $\beta^{(x)} = 0$ for all transient noise NMF coefficients.

After estimating $\mathbf{v}_t$ using (5), the speech DFT magnitudes are estimated using a gain function that is obtained by dividing the speech NMF approximation $\mathbf{B}^{(s)}\hat{\mathbf{v}}_t^{(s)}$ by the input NMF approximation $\mathbf{B}\hat{\mathbf{v}}_t$.

3.2. Enhancement Strategy using BNMF

Similarly to [9], we assume that y_{kt} is sampled from a Poisson distribution whose mean is given by the kt-th element of $\mathbf{BV}$. This is equivalent to $y_{kt} = \sum_i z_{kit}$ in which z_{kit} are a set of Poisson-distributed hidden variables with mean values given by $b_{ki}v_{it}$. We further assume that b and v are random variables with gamma prior distributions, which is a common choice for nonnegative data.

The first step of the enhancement algorithm is to apply BNMF on $\mathbf{y}_t$ to obtain the posterior distribution of the NMF coefficient vector $\mathbf{v}_t$, i.e., $f(\mathbf{v}_t \mid \mathbf{y}_t, \mathbf{B})$. Since obtaining the exact posterior distribution is intractable, a variational Bayes (VB) approach has been proposed in [9] that approximates $f(v_{it} \mid \mathbf{y}_t, \mathbf{B})$ by a gamma distribution. The utilized VB approach is an iterative scheme that maximizes a lower bound on the marginal log-likelihood of the observations.

In the second step of the enhancement phase, the mean square error $E((s_{kt} - \hat{s}_{kt})^2)$ is minimized to obtain an estimate of the speech DFT magnitude $\hat{s}_{kt}$. The MMSE estimate is given by the mean of the posterior distribution of s_{kt} and can be derived similarly to [9] to obtain (6).

In order to use the temporal dependencies of the speech and the noise signals, we assume that the mean of the NMF coefficients v_{it} can be modeled using an autoregressive model, i.e.,

$$E(v_{it}) = \sum_{m=1}^{t-1} (1-\alpha)\,\alpha^{m-1}\hat{v}_{i(t-m)},$$

$$f(v_{it}) = \text{gamma}(\phi_i, E(v_{it})/\phi_i), \tag{7}$$

$$\hat{s}_{kt} = E(s_{kt} \mid \mathbf{y}_t) = \frac{\sum_i e^{E\left(\log b_{ki}^{(s)} + \log v_{it}^{(s)} \mid \mathbf{y}_t\right)}}{\sum_i e^{E\left(\log b_{ki}^{(s)} + \log v_{it}^{(s)} \mid \mathbf{y}_t\right)} + \sum_i e^{E\left(\log b_{ki}^{(n)} + \log v_{it}^{(n)} \mid \mathbf{y}_t\right)} + \sum_i e^{E\left(\log b_{ki}^{(x)} + \log v_{it}^{(x)} \mid \mathbf{y}_t\right)}} y_{kt}. \tag{6}$$

where $\alpha < 1$ specifies the AR parameters, $\hat{v}_{i(t-m)}$ is the mean of the posterior distribution $f(v_{i(t-m)} \mid \mathbf{y}_{t-m}, \mathbf{B})$, and gamma$(a, b)$ is a gamma distribution with a and b denoting the shape and scale parameters, respectively. Note that the first line in (7) is equivalent to $E(v_{it}) = \alpha E(v_{i(t-1)}) + (1-\alpha)\hat{v}_{i(t-1)}$. In addition, ϕ_i is a signal-specific parameter which is related to the stationarity of the signal. We use a single shape parameter $\phi^{(s)}$ for all speech NMF coefficients. Similarly, two different shape parameters $\phi^{(n)}$ and $\phi^{(x)}$ are used for the NMF coefficients corresponding to the background and transient noises. An initial value for these parameters can be obtained by examining the posterior distributions of the NMF coefficients of the training data but further tuning might be required to obtain the best performance [9].

As mentioned earlier, when no transient noise is present, setting a flat prior for the speech NMF coefficients, which corresponds to setting the shape parameter $\phi^{(s)}$ to a very small positive number, yields the best performance [9]. In this case, the corresponding NMF coefficients will take maximum likelihood (ML) estimates. However, when transient noise that is more non-stationary than speech is present, exploiting the temporal correlations in the speech signal is also beneficial (cf. Section 4). For this purpose, we use a larger value for $\phi^{(s)}$. In our experiments, we set $\phi^{(x)}$ to a very small positive number (corresponding to a flat prior), while $\phi^{(s)}$ was set to 0.25.

3.3. Dictionary Learning

In the previous subsections we assumed that the dictionaries $\mathbf{B}^{(s)}$, $\mathbf{B}^{(x)}$, and $\mathbf{B}^{(n)}$ are given. In the following we discuss how these dictionaries are learnt and updated.

First, the *speech dictionary* $\mathbf{B}^{(s)}$ is learnt using a large speech corpus with different speakers to make it universal and speaker-independent. Let us denote the magnitude spectrogram of the training data by $\overline{\mathbf{S}}$. We use PLCA or BNMF to obtain $\overline{\mathbf{S}} \approx \mathbf{B}^{(s)}\overline{\mathbf{V}}^{(s)}$. In the case of BNMF, we use a variational Bayes approach to obtain the posterior distribution of $\mathbf{B}^{(s)}$. The dictionary $\mathbf{B}^{(s)}$ (or the posterior distribution of $\mathbf{B}^{(s)}$) is then held fixed during the enhancement.

The *transient noise dictionary* $\mathbf{B}^{(x)}$ is learnt using samples from each individual transient noise type and is therefore noise-dependent. The learning mechanism is similar to that of $\mathbf{B}^{(s)}$. During the enhancement phase, we assume that we know the noise type and we select the proper dictionary to enhance the noisy signal. Using a noise-independent dictionary for the transient noise, or updating $\mathbf{B}^{(x)}$ online remains an open question for further research.

Unlike $\mathbf{B}^{(s)}$ and $\mathbf{B}^{(x)}$, the *background noise dictionary* $\mathbf{B}^{(n)}$ is learnt online using the noisy observations, cf. Fig. 1. This dictionary is updated such that it captures the structure of the remaining background noise that is assumed to be more stationary than the speech signal. Our online dictionary learning scheme is based on a sliding window concept [9]. Hence, recent noisy frames are first stored in a buffer. Then, a number of these frames with lowest energies are chosen and collected in a matrix denoted by $\mathbf{N}$. Moreover, to ensure that the background noise dictionary is updated smoothly over time, the current estimate of $\mathbf{B}^{(n)}$ is used to construct an informative prior to be used in the new estimation problem. In the BNMF approach, the prior distribution for the noise dictionary is updated similarly to the first line of (7). Then, BNMF is applied to $\mathbf{N}$ and using the variational Bayes method the posterior distribution of $\mathbf{B}^{(n)}$ is computed while the other dictionaries are held fixed. A similar procedure can be adapted for the PLCA method. More details about the online learning approach can be found in [9]. Moreover, a Matlab implementation of this algorithm is available in [13].

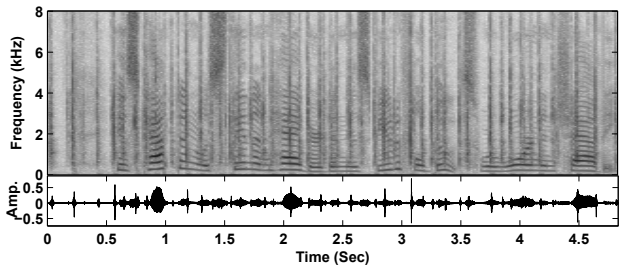
4. EXPERIMENTAL RESULTS

We evaluated the proposed NMF-based transient noise reduction algorithms for keyboard typing and machinegun noises taken from the Sound Ideas and NOISEX databases [14, 15], respectively. The signals were divided into two disjoint training and testing sets. The dictionary size was set experimentally to obtain the best results. For the BNMF models, we learned 10 and 5 elements per dictionary for keyboard and machinegun noise types, respectively, while for the PLCA models we learned 30 elements for both noise types. For both BNMF and PLCA speech models, 60 dictionary elements were learnt for both BNMF and PLCA speech models using the training sentences from the TIMIT database [16]. The test speech signal was obtained by concatenating 20 sentences uttered by different speakers from the core test set of the TIMIT database. All signals were downsampled to 16 kHz, and the signal synthesis was performed using the overlap-add procedure using a frame length of 512 samples with half-overlapped Hann windows.

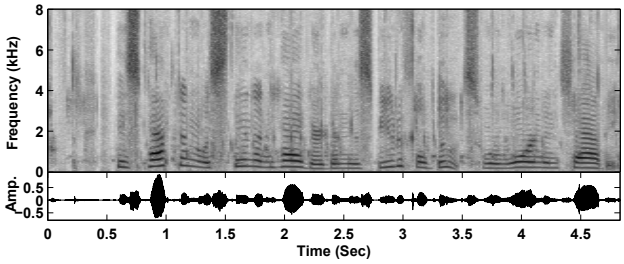
First, the enhancement performance is assessed by comparing the spectrograms of the noisy and the enhanced speech signals. Fig. 2 shows the spectrograms and waveforms of the noisy, clean, and enhanced speech signals uttered by a female speaker. In this example, no stationary background noise is present and keyboard typing noise is added to the clean speech signal at 5 dB signal-to-noise ratio, which is denoted by $\text{SNR}_{\text{tr}} = 5$ dB. The enhanced signals using the BNMF and the PLCA approaches are shown in Fig. 2b and Fig. 2c, respectively. The figure shows that both methods have significantly reduced the interfering noise while the speech signal remains highly undistorted. Considering the BSS-Eval metrics [17], in this particular example the source to interference ratio (SIR) and source to artifact ratio (SAR) are around 20 dB and 10 dB, respectively. Moreover, computing the perceptual evaluation of speech quality (PESQ) [1] shows a significant quality improvement (around 1 MOS in this experiment). Comparing the two enhancement approaches, we can see that the BNMF method results in a higher interference suppression and less speech distortion (e.g., compare the plots around 1.3 second).

Second, to have a more thorough comparison of the PLCA and BNMF methods, we considered noisy signals at different SNR_{tr} for both keyboard and machinegun noises (without stationary background noise). Fig. 3 shows the SIR and PESQ improvement for different transient noise levels. This figure compares the BNMF and PLCA approaches with and without temporal continuity. As can be seen, the standard PLCA and BNMF approaches yield good performance (especially BNMF with a PESQ improvement of around 0.5 MOS). By additionally using the speech temporal dependencies (as explained earlier, the priors corresponding to the transient noise are set to be flat) a large improvement is observed for both approaches. In addition, these experimental results show that BNMF leads to a better performance than PLCA.

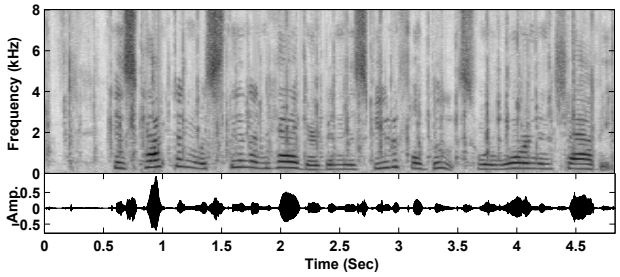
Finally, we consider a noise reduction problem where both transient and background noises are present. As the BNMF scheme outperformed the PLCA approach in the previous experiment, we only evaluate the performance of the BNMF method in this experiment. Here, the noisy observation was a mixture of speech, keyboard typing noise (level determined by SNR_{tr}), and white Gaussian noise (level determined by SNR_{bg}). We used online BNMF to learn 5 dictionary elements for the background noise while the other online learning parameters were set similarly to [9]. Table 1 shows the SIR and PESQ improvement for different combinations of SNR_{bg} and SNR_{tr}. The table additionally shows the PESQ measure evaluated for the input noisy signals. The results show a significant amount of



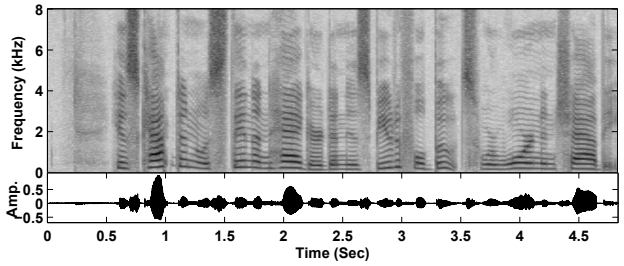
(a) Noisy signal



(b) Enhanced signal using BNMF



(c) Enhanced signal using PLCA



(d) Clean speech signal

Fig. 2: Spectrograms and signal waveforms for a speech signal uttered by a female speaker. The noisy signal is obtained by adding the keyboard typing noise at 5 dB signal-to-noise ratio. The figure shows that the noise level is significantly reduced using both PLCA and BNMF schemes while the BNMF method results in a higher interference suppression and less speech distortion, e.g., compare the plots around 1.3 second.

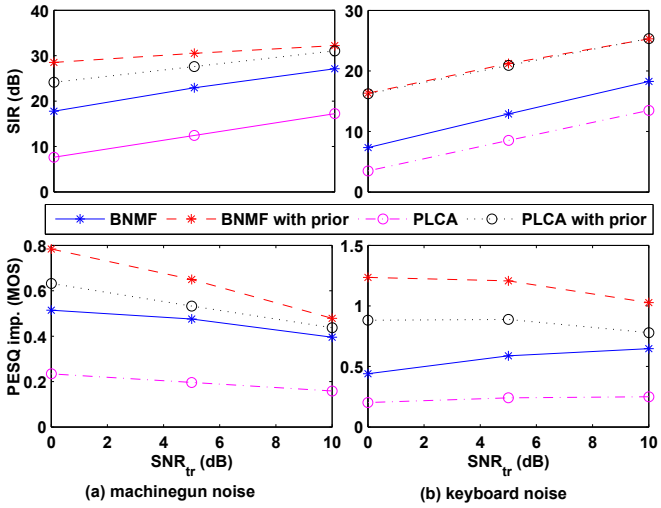


Fig. 3: SIR and PESQ improvement using the BNMF and PLCA methods for machinegun noise (left) and keyboard typing noise (right) for different noise levels.

Table 1: Performance measures in the presence of both keyboard typing noise and white Gaussian noise for the BNMF method where the non-transient background noise dictionary is learnt online from the noisy observation.

(a) Enhancement result as a function of SNR_{bg} with SNR_{tr}=10 dB.

SNR_{bg} (dB)	0	5	10	15	20	30
SIR (dB)	14.9	21.8	24.5	25.5	26	26
PESQ imp. (MOS)	0.84	0.80	0.76	0.80	0.81	0.78
Input PESQ (MOS)	1.4	1.7	1.8	1.9	1.9	1.9

(b) Enhancement result as a function of SNR_{tr} with SNR_{bg}=10 dB.

SNR_{tr} (dB)	0	5	10	15	20	30
SIR (dB)	18	22	24.5	26	26.6	27
PESQ imp. (MOS)	0.97	0.93	0.76	0.61	0.55	0.51
Input PESQ (MOS)	1.2	1.5	1.8	2	2.2	2.2

noise reduction and speech quality improvement at different noise levels. As can be seen, when the level of one of the interfering noises becomes much smaller than the other one (e.g., $SNR_{tr} \in [15, 20, 30]$ in Table 1b), the measures converge to some stationary values which is mainly determined by the dominant noise signal.

5. CONCLUSION

This paper investigated the application of Bayesian NMF (BNMF) and PLCA methods for transient noise reduction. Using the temporal dependencies of the speech and the background noise, we designed informative priors that significantly improved the noise reduction performance. In our experiments, the BNMF-based approach outperformed the PLCA-based approach both in terms of PESQ and SIR. Additionally, we used online NMF to learn the dictionary of the background noise and used it to enhance a noisy signal contaminated with both transient noise and white Gaussian noise. Our experiments show that the proposed methods can be reliably used to enhance the noisy signal and achieve promising performance.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Press, 2007.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-domain based single-microphone noise reduction for speech enhancement - a survey of the state of the art," Synthesis Lectures on Speech and Audio Processing, Morgan & Claypool Publishers, Jan. 2013.

[3] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[4] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*, 4th ed. Wiley, 2008.

[5] A. Sugiyama and R. Miyahara, "Tapping-noise suppression with magnitude-weighted phase-based detection," in *Proc. IEEE Workshop Applications of Signal Process. Audio Acoust. (WASPAA)*, Oct. 2013, pp. 1–4.

[6] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Supervised graph-based processing for sequential transient interference suppression," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 9, pp. 2528–2538, 2012.

[7] 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.

[8] 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)*, May 2013, pp. 873–877.

[9] ——, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[10] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[11] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 1, pp. 132–144, Jan. 2013.

[12] P. Smaragdis, B. Raj, and M. V. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Advances in Models for Acoustic Process. Workshop, NIPS*. MIT Press, 2006.

[13] N. Mohammadiha, "Matlab codes for speech enhancement using NMF." [Online]. Available: http://www.sigproc.uni-oldenburg.de/63074.html

[14] B. Nimens *et al.*, "Sound ideas: sound effects collection," ser. 6000, http://www.sound-ideas.com/6000.html.

[15] A. Varga and H. J. M. 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.

[16] 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.

[17] 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.