International Journal of Multimedia and Ubiquitous Engineering
Vol.9, No.7 (2014), pp.73-86
http://dx.doi.org/10.14257/ijmue.2014.9.7.07

Speech Intelligibility Enhancement Using Convolutive Non-negative Matrix Factorization with Noise Prior

Jian Zhou[1], Xianyong Fang[1], Liang Tao[1] and Li Zhao[2]

[1]*Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, 230601 Hefei, China*
[2]*Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, 210096 Nanjing, China*
230099140@seu.edu.c

Abstract

We propose a convolutive non-negative matrix factorization method to improve the intelligibility of speech signal in the context of adverse noise environment. The noise bases are prior learned with Non-negative Matrix Factorization (NMF) algorithm. A modified convolutive NMF with sparse constraint is then derived to extract speech bases from noisy speech. The divergence function is selected as an objective function to get a multiplicative update of speech base and its corresponding weight. The weights of prior learned noise bases are also updated in the update rule. Listening experiments are conducted to assess the intelligibility performance of speech synthesized using the proposed algorithm. Experimental results indicate that the proposed method is very effective to improve the intelligibility of the noisy speech in various noise contexts and it outperforms conventional algorithms.

Keywords: *Speech intelligibility, Speech enhancement, Non-negative Matrix Factorization, Speech base learning*

1. Introduction

During the last few decades, many speech enhancement algorithms have been proposed for processing of noisy speech in stationary or non-stationary environment. And much progress has been made in the aspect of improving speech quality (*e.g.*, SNR improvements or comfort of the enhanced speech) by these algorithms. However, considerably smaller progress has been made in designing algorithms that can improve speech intelligibility [1].

Quality and intelligibility are two dimensions used to measure the perception of a speech signal. The quality is a subjective measure which reflects on individual preferences of listeners, it is relevant to the signal-to-noise ratio (SNR) improvements or comfort of the listener. However, intelligibility is an objective measure which predicts the percentage of words that can be correctly identified by listeners. It has strong relation to the understanding of the underlying message or content of speech signal [1].

Generally, improving the intelligibility of noisy speech rather than the quality is much more important in adverse environment since semantic information retrieval becomes the dominating purpose of speech communication. In [2], Kim proposed an algorithm that can be optimized for a specific acoustic environment and improve speech intelligibility. The proposed method decomposes the input signal into time-frequency (T-F) units and makes binary decisions, based on a Bayesian classifier, as to whether each T-F unit is dominated by the target signal or the noise masker. However, this method may fail in lower SNR

ISSN: 1975-0080 IJMUE

environment wherein the speech energy dominated T-F unit is easily classified as the noise dominated unit by the supervised learning approach.

Individual sources identification from complex mixtures in adverse noisy context is a powerful ability of human beings. For example, if a violin section plays in unisono, all the sounds arriving from the same musical instrument are perceived as a single source. This remarkable ability is attributed to the human auditory perceptual system [3,4]. According to perceptual principles, a collection of sensory elements likely arise from the same source are grouped and perceived while others are discarded. These sensory elements can be T-F units where the object speech dominates in the joint time-frequency domain [5].

In order to understanding the perception principle of the brain, Lee and Seung proposed a parts-based presentation which is confirmed by psychological and physiological evidence [6]. For example, the lateral occipital complex (LOC), a cortical region critical for human object recognition, has been shown to primarily code the shape, rather than the surface properties, of an object [7]. Certain computational theories such as computational auditory scene analysis (CASA) also rely on such representations [3, 8]. As an emerging approach for data analysis, NMF has been successfully used in such applications as music transcription, blind source separation and *et al.* [9-12].

In parts-based representation, linear combinations of basis vectors is forced to be non-subtract, since negative basis components are physically meaningless in such applications as image and audio processing. The non-negative matrix factorization (NMF) proposed by Lee and Seung is a parts-based algorithm where basis decomposition is confined to a non-negative space, *e.g.*, NMF approximately decomposes a non-negative matrix $\mathbf{V}_{m\times n}$ into a product of two non-negative matrix $\mathbf{B}$ and $\mathbf{H}$. The basis vectors $\mathbf{B}$ describe the spectral characters of the components, whereas, their weights $\mathbf{H}$ provide their temporal evolution. In audio processing, $\mathbf{V}$ is usually the joint time-frequency (magnitude or power) spectrogram with m frames and n frequency bins.

In some tasks such as speech signal processing, a frequency component of speech often spans multiple columns of $\mathbf{V}$. However, conventional NMF is ignoring potential dependencies across successive columns of $\mathbf{V}$. In order to capture the temporal dependency of the frequency patterns within the signal, Smaragdis proposed a convolutive NMF (cNMF) which is utilized successfully in separating monophonic mixtures of known speakers [13].

The drawback of the cNMF algorithm proposed in [13] is that speech bases of each speakers must be got prior in the separating task. This is, however, usually impossible in single-channel speech enhancement since the speech signal of specific speaker cannot be captured quietly in the adverse noise environment.

To tackle this problem, we propose a novel convolutive NMF algorithm to conduct speech intelligibility enhancement. Firstly, the noise bases are prior learned using the conventional cNMF. New update rules are then derived using the divergence object function with sparseness constraint. Estimated clean speech bases are learned using the proposed algorithm thereafter. The estimated speech is reconstructed using the learned phoneme bases and its corresponding weights. The object function is selected according to perceptual principle. The algorithm allows making the spectrum of the residual noise similar to that of the speech signal in the given frame. It expects to comprise most speech sensory parts, leading to the enhanced speech more intelligible.

The remainder of this paper is organized as follows. The next section describes non-negative matrix factorization and its convolutive version. Section 3 presents a novel convolutive NMF algorithm for the purpose of speech intelligibility enhancement. In section 4 we apply the proposed convolutive NMF to do speech intelligibility enhancement. The

performance is also compared with other classic speech enhancement algorithms. The paper concludes with section 5.

2. Convolutive Non-negative Matrix Factorization

Given an $M \times N$ non-negative matrix $\mathbf{V} \in \mathbb{R}^{M \times N}$, the goal of NMF is to find non-negative matrix $\mathbf{B} \in \mathbb{R}^{M \times R}$ and $\mathbf{H} \in \mathbb{R}^{R \times N}$, such that $\mathbf{V} \approx \mathbf{BH}$. R is the rank of the factorization and generally chosen to be smaller than M (or N), or akin to $(M + N)R < MN$, which result in the extraction of some latent features whilst reducing some redundancies in the input data. To find $\mathbf{B}$ and $\mathbf{H}$ such that the reconstruction error is minimized, two error cost functions have been proposed in [14]. One cost function is the squared Euclidean distance which is defined as $L(\mathbf{V}, \mathbf{B}, \mathbf{H}) = 1/2 \| \mathbf{V} - \mathbf{BH} \|_F^2$ and the other is the divergence function which is defined as

$$D(\mathbf{V}, \mathbf{B}, \mathbf{H}) = \| \mathbf{V} \otimes ln(\mathbf{V} \,\%\, \mathbf{BH}) - \mathbf{V} + \mathbf{BH} \|_1, \tag{1}$$

where $\otimes$ and $\%$ denote the element-wise product and division, respectively, $\| \|_1$ denotes l_1-norm. The divergence is lower bounded by zero, which is attained if and only if $\mathbf{V} = \mathbf{BH}$. The divergence reduces to Kullback-Leibler divergence when $\sum_{i,j} \mathbf{V}_{ij} = \sum_{i,j} (\mathbf{BH})_{ij} = 1$. The use of divergence as cost function is motivated by the fact that the divergence is less sensitive to large-energy observations than the Euclidean distance [10]. This enables the use of power spectrogram as the observation.

In order to find a local minimum value of the cost function, Lee and Seung proposed a very appealing multiplicative update algorithm described as follows [14],

$$\mathbf{H}^{q+1} = \mathbf{H}^{q} \otimes ((\mathbf{B}^{q})^T \mathbf{V}) \,\%\, ((\mathbf{B}^{q})^T \mathbf{B}^{q} \mathbf{H}^{q}), \tag{2}$$

$$\mathbf{B}^{q+1} = \mathbf{B}^{q} \otimes (\mathbf{V} (\mathbf{H}^{q+1})^T \,\%\, (\mathbf{B}^{q} \mathbf{H}^{q+1} (\mathbf{H}^{q+1})^T), \tag{3}$$

where T is the matrix transpose operator, and q is the iteration index.

In the aforementioned standard NMF, each object is composed of a single spectrum which is calculated over the duration of the spectrogram frame. It is however a weak model since it does not take into account the relative positions of each spectrum thereby discarding temporal information. For example, the frequency spectrum component of a speech object (for example, phoneme) often spans several successive frames.

In order to capture this time-varying auditory object of speech, a convolution version of NMF is proposed by Smaragdis in [13] as follows,

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{B}(t) \overset{t\rightarrow}{\mathbf{H}}, \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{M \times N}$ is the input we wish to decompose, $\mathbf{B}(t) \in \mathbb{R}^{M \times R}$ and $\mathbf{H} \in \mathbb{R}^{R \times N}$ are the bases and weights matrix respectively. The function $\overset{t\rightarrow}{(.)}$ denotes a column shift operator that moves its arguments by t spots to the right; as each column is shift off to the right the leftmost columns are zero padded. And consequently $\overset{t\leftarrow}{(.)}$ shifts to the left.

From Eq.(4), one can find that the convolution NMF is essentially a summation of convolution operations between corresponding elements from a set of two-dimensional

bases $\mathbf{B}$ and a set of weights $\mathbf{H}$. When $T = 1$, it degenerates to the standard NMF. All the j th column of $\mathbf{B}(t), t = 0,1,...,T-1$ consists of a two-dimensional auditory object which we also refer to as a basis. With this convolutive NMF, the temporal continuity possessed by many audio signals can be captured and represented more effectively in the joint time-frequency domain, especially for speech signals whose frequencies vary with time [15].

Multiplicative updates rules are also derived in [13] to conduct convolution NMF based on extend Kullback-Leibler divergence, which can be rewritten in matrix form as:

$$\mathbf{H}^{q+1} = \mathbf{H}^{q} \otimes ((\mathbf{B}^{q}(p))^{T} \mathbf{V} \;\%\; \overset{p\leftarrow}{\mathbf{\Lambda}}^{q}) \;\%\; ((\mathbf{B}^{q}(p))^{T} \mathbf{\Xi})), \tag{5}$$

$$\mathbf{B}^{q+1}(p) = \mathbf{B}^{q}(p) \otimes (((\mathbf{V} \;\%\; \mathbf{\Lambda}^{q})(\overset{p\rightarrow}{\mathbf{H}}^{q+1})^{T}) \;\%\; (\mathbf{\Xi}(\overset{p\rightarrow}{\mathbf{H}}^{q+1})^{T})), \tag{6}$$

where $\mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{B}(t) \overset{t\rightarrow}{\mathbf{H}}$ and $\mathbf{\Xi}$ is an $M \times N$ matrix whose elements are all set to unity.

3. A Novel Convolutive NMF Algorithm for Speech Intelligibility Enhancement

In this section, we first give an introduction to the signal model. The proposed novel convolutive NMF algorithm is then derived and described in detail.

3.1. Signal Model

We assume that the noisy speech $v(i)$ at sampling time index i consists of speech $s(i)$ and additive noise $n(i)$. For joint time-frequency analysis of $v(i)$, we apply the K -point STFT, that is,

$$V(\lambda, k) = \sum_{\mu=0}^{L-1} v(\lambda R + \mu) h(\mu) exp(-\frac{j 2\pi k \mu}{K}), \tag{7}$$

where $\lambda \in Z$ is the sub-sampled time index, $k = 0,1,\ldots,K-1$ is the frequency bin, and L is the window length. The quantity R is the number of samples that successive frames are shifted and $h(\mu)$ is a unit-energy window function, that is $\sum_{\mu=0}^{L-1} h^{2}(\mu) = 1$.

From the linearity of Eq.(7), we have that

$$V(\lambda, k) = S(\lambda, k) + N(\lambda, k), \tag{8}$$

where $S(\lambda, k)$ and $N(\lambda, k)$ are the STFT coefficients of speech $s(i)$ and additive noise $n(i)$, respectively. We further assume that $s(i)$ and $n(i)$ are zero mean and statistically independent, which leads to a power relation, where the noise is additive, that is,

$$|V(\lambda, k)|^{2} = |S(\lambda, k)|^{2} + |N(\lambda, k)|^{2} \tag{9}$$

Eq. (9) can be rewritten in matrix form as follow:

$$\mathbf{V} = \mathbf{S} + \mathbf{N} \tag{10}$$

3.2. Derivation of the Proposed Algorithm

Since matrix $\mathbf{V}$, $\mathbf{S}$ and $\mathbf{N}$ are non-negative, we can decompose $\mathbf{V}$ with convolutive NMF,

$$\mathbf{V} \approx \mathbf{\Lambda} = \sum_{t=0}^{T-1} \mathbf{B}(t) \overrightarrow{\mathbf{H}}^{t} = \sum_{t=0}^{T-1} [\mathbf{B}^{s}(t) \quad \mathbf{B}^{\hat{0}}(t)] \begin{bmatrix} \overrightarrow{\mathbf{H}}^{t}{}^{s} \\ \overrightarrow{\mathbf{H}}^{t}{}^{\hat{0}} \end{bmatrix} \tag{11}$$

The fact that some of the eigenvalues of the clean signal may be zero indicates that the energy of the clean signal vector is distributed among a subset of its coordinates, e.g, the signal is confined to a subspace of the noisy Euclidean space, while the uncorrelated noise fills in the entire vector space of the noisy signal. In order to utilize this character in the convolutive NMF, we give an addition of the sparseness constraint on $\mathbf{H}^{s}$. Combining Eq.(1) with sparseness constraint on $\mathbf{H}^{s}$ results in the following objective function,

$$G(\mathbf{V}, \mathbf{\Lambda}, \mathbf{H}^{s}) = \| \mathbf{V} \otimes ln(\mathbf{V} \,\%\, \mathbf{\Lambda}) - \mathbf{V} + \mathbf{\Lambda} \|_{1} + \lambda \sum_{j,k} H^{s}_{j,k}, \tag{12}$$

where the left term of the objective function corresponds to the conventional convolutive NMF, the right term is an additional constraint on $\mathbf{H}^{s}$ that enforces sparsity by minimizing the L_1-norm of its elements. The parameter λ controls the tradeoff between sparseness and accurate reconstruction.

The extended objective function of Eq.(12) introduces a scaling problem: The right term is a strictly increasing function of its argument, so it is possible that the object can decreased by scaling $\mathbf{B}^{s}$ up and $\mathbf{H}^{s}$ down. This situation does not alter the left term but will cause the right term to decrease, resulting in the elements of $\mathbf{B}(t)^{s}$ growing without bound and $\mathbf{H}^{s}$ tends toward zero. To avoid this scaling problem, normalizing $\mathbf{B}(t)^{s}$ is performed for each object matrix $\mathbf{B}_{j}(t)^{s}$ to be scaled to the unit L_2-norm,

$$\overline{\mathbf{B}}^{s}_{j} = \frac{\mathbf{B}^{s}_{j}}{\| \mathbf{B}^{s}_{j} \|}, \; j = 1, \ldots, R, \tag{13}$$

where the matrix $\mathbf{B}^{s}_{j}$ is constructed from the j th of $\mathbf{B}(t)^{s}$ at each time step. Since the noise can be estimated for example by VAD algorithm from the noisy speech, the noise bases $\mathbf{B}^{n}$ can be learned from these roughly noise signal. The remainder problem need to tackle is to estimate $\mathbf{B}^{s}$, $\mathbf{H}^{s}$ and $\mathbf{H}^{n}$. When this problem is approached, the enhanced speech can be reconstructed by,

$$\mathbf{S} = \sum_{t=0}^{T-1} \mathbf{B}^{s}(t) \overrightarrow{\mathbf{H}}^{t}{}^{s} \tag{14}$$

We first derive the update rules for $\mathbf{B}^{s}$. According to Eq.(11), we have the derivative $\Lambda_{i,j}$ with respect to $B^{s}_{m,n}(t)$,

$$\frac{\partial \Lambda_{i,k}}{\partial B^{s}_{m,n}(t)} = \frac{\| \mathbf{B}^{s}_{n} \| \overrightarrow{h}^{t}{}^{s}_{nk} - B_{mn}(t) \overrightarrow{h}^{t}{}^{s}_{nk} \overline{B}^{s}_{mn}(t)}{\| \mathbf{B}^{s}_{n} \|^{2}}, \tag{15}$$

where $\overline{B}^{s}_{mn}(t) = \partial \| \mathbf{B}_{j} \| / \partial B^{s}_{mn}(t)$.

Now, we have the derivative of G with respective to $B^{s}_{m,n}(t)$,

$$\frac{\partial G(\mathbf{V},\mathbf{\Lambda},\mathbf{H}^s)}{\partial B^s_{m,n}(t)} = \sum_k \frac{\{\| \mathbf{B}^s_n\| \overrightarrow{h^s_{nk}}^t - B_{mn}(t)\overrightarrow{h^s_{nk}}^t \overline{B}^s_{mn}(t)\}(1-\frac{V_{m,k}}{\Lambda_{m,k}})}{\| \mathbf{B}^s_n\|^2} \tag{16}$$

The gradient descent update for H can be derived as,

$$B_{m,n}(t) = \eta_{B^s_{mn}(t)} \frac{\partial G}{\partial B^s_{m,n}(t)} + B_{m,n}(t) \tag{17}$$

Let the element-wise step size be

$$\eta_{B^s_{mn}(t)} = \frac{B^s_{mn}(t)}{\sum_{k=1}^{N} \frac{\overrightarrow{h^s_{nk}}^t V_{m,k}\| \mathbf{B}_n\| + B^s_{mn}(t)\overrightarrow{h^s_{nk}}^t \overline{B}^s_{mn}(t)}{\| \mathbf{B}\|^2 \Lambda_{m,k}}}, \tag{18}$$

then substitute Eq.(18) into Eq. (17), we get the following multiplicative update rules of $B_{m,n}(t)$ as

$$B_{m,n}(t) = \frac{\sum_{k=1}^{N} \overrightarrow{h^s_{nk}}^t (1+\overline{B}^s_{mn}(t)V_{m,k}\overline{B}^s_{mn}(t)\Lambda^{-1}_{m,k})B_{m,n}(t)}{\sum_{k=1}^{N} \overrightarrow{h^s_{nk}}^t (V_{m,k}\Lambda^{-1}_{m,k} + \overline{B}^s_{mn}(t)\overline{B}^s_{mn}(t))}, \tag{19}$$

where $t = 0,\ldots,T-1$.Similarly, we derive a new update for $\mathbf{H}^s$,

$$h^s_{j,k} = h^s_{j,k} + \eta_{h^s_{j,k}} \frac{\partial G}{\partial h^s_{j,k}}. \tag{20}$$

Using the gradient

$$\frac{\partial G}{\partial h^s_{j,k}} = \lambda + \sum_{i=1}^{M} (1 - V_{i,k} / \overleftarrow{\Lambda_{i,k}}^t)\overline{B}^s_{m,n}(t). \tag{21}$$

Setting the learning rate to

$$\eta_{h^s_{j,k}} = \frac{h^s_{j,k}}{\lambda + \sum_{i=1}^{M} \overline{B}^s_{m,n}(t)}. \tag{22}$$

We get the following multiplicative rules for $\mathbf{H}^s$,

$$h^s_{j,k} = \frac{\sum_{i=1}^{M} \overline{B}^s_{m,n}(t)(V_{i,k} / \overleftarrow{\Lambda_{i,k}}^t)h^s_{j,k}}{\lambda + \sum_{i=1}^{M} \overline{B}^s_{m,n}(t)}. \tag{23}$$

The update rules for $\mathbf{H}^n$ can be derived similarly as follows,

$$h^n_{j,k} = \frac{\sum_{i=1}^{M} \overline{B}^{\hat{o}}_{m,n}(t)(V_{i,k} / \overleftarrow{\Lambda_{i,k}}^t)h^n_{j,k}}{\sum_{i=1}^{M} \overline{B}^{\hat{o}}_{m,n}(t)} \tag{24}$$

4. Numerical Simulation Experiment

In this section, we study numerically the performance of the proposed algorithm in the context of speech intelligibility enhancement and perform comparisons to the conventional enhancement algorithms.

4.1. Corpus and Power Spectra

In order to evaluate the performance of the proposed method, 30 clean utterances obtained from IEEE speech database were used in the experiments [16]. Half of the utterances are from male speakers and half are from female speakers. Three types of noise recordings including Gaussian white noise, babble noise and F16 fighter jet noise taken from NOISEX-92 database were used as noise maskers [17]. A noise segment of the same length as the speech signal was randomly cut out of the noise recordings and appropriately scaled to reach the desired SNR level. Power spectrums of noisy speech and noise signals are computed respectively.

A fixed 32-ms frame size was used with 50% overlap between frames. The discrete Fourier transform (DFT) was applied on each frame, the length of the DFT is equal to the frame size which has length of 256 sample points. Only positive frequencies are retained and phases were discarded by taking the square values of the DFT spectra, resulting in power spectrogram matrix $V_{k,l}$ corresponding to noisy signal and $N_{k,l}$ corresponding to pure noise signal where k is the discrete frequency index and l is the frame index.

4.2. Noise Base Training

The scNMF algorithm proposed by Smaragdis was used to find noise basis. The pure noise signal which was also contained in the noisy speech was processed prior similarly as the noisy speech to get its power spectra. The noise bases were got after 50 iterations. The sparse constraint parameter λ is set to zero which indicates that there is no sparseness constraint. The number of objects in noise is set as the same as that in the clean speech. Figure 1 shows 10 bases which trained in the F16 fighter jet noise. In Figure 1, the top panel plots the spectrum of F16 noise and the bottom panel plots 10 bases which are learned from successive 23 frames.

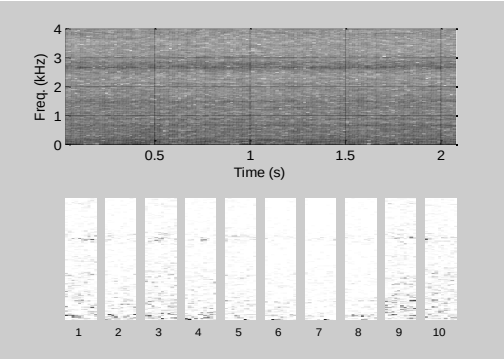

Figure 1. F16 Fighter Jet Noise Spectrum and its 10 bases Learned using scNMF Proposed in [13]. The Top Panel Plots the Spectrum of F16 Noise and the Bottom Panel Plots 10 bases which are Learned from Successive 23 Frames

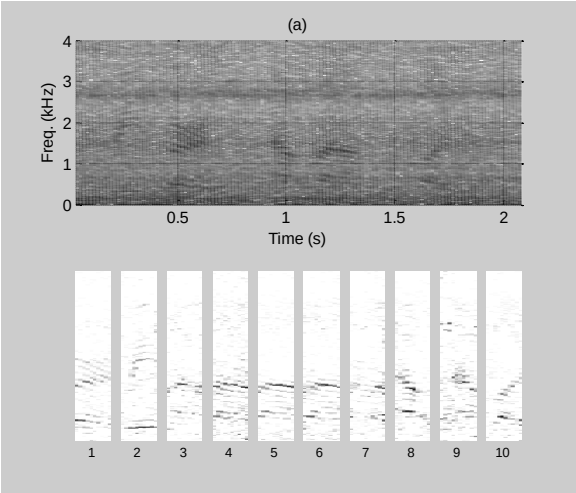

Figure 2. Noisy Speech and Speech bases at SNR of -5 dB

The top panel plots the noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. The bottom panel plots 10 phoneme bases learned using the proposed convolutive NMF algorithm.

4.3. Speech Intelligibility Enhancement

We use the algorithm described in section 3 to estimate the speech phoneme patterns from the noisy signal. The noise bases trained in the above section were used as fixed noise bases. The objective function Eq.(12) was used with sparseness constraint $\lambda = 0.15$. Eq. (19), (23) and (24) were used to update $w_{m,n}(t)$, $h^{s}_{j,k}$ and $h^{n}_{j,k}$ respectively. We did the update rules with 40 iterations to get a local minimum of the objective function. We then used Eq. (14) to reconstruct the enhanced speech where $\mathbf{B}^{s}(t)$ consists of the estimated phoneme bases and $\mathbf{H}^{s}$ is the corresponded weights matrix.

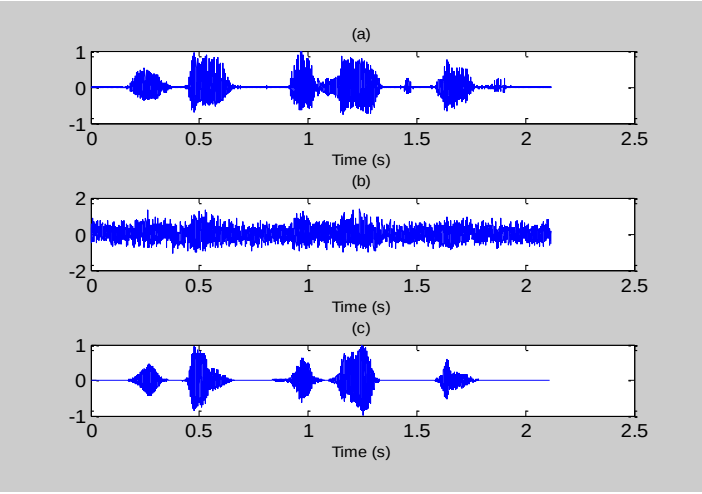

Figure 3. Reconstructed Speech with Learned Speech Bases at SNR of -5 dB

Figure 3 (a) shows the original clean speech, Figure 3 (b) shows the noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. Figure 3 (c) shows the reconstructed speech using the learned phoneme bases and their weights. Figure 2 shows the base phoneme bases extracted from noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. In Figure 2, The top panel plots the noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. The bottom panel plot 10 phoneme bases learned using the proposed convolutive NMF algorithm. As can be seen from Figure 2, the speech bases learned from the noisy speech captured the phoneme characteristic. Compared with the noise bases in Figure 1, each base is the frequency spectrum component of a speech object, containing little of noise signal.

Figure 3 shows the enhanced speech reconstructed using the proposed algorithm from noisy speech which contaminated by F16 fighter jet noise. Figure 3 (a) shows the original clean speech signal, Figure 3 (b) shows the noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. Figure 3(c) shows the reconstructed speech using the learned phoneme bases and their weights. From Figure 3, we can find that the noise has been cleaned mostly while the speech components have been retained. Although some speech components have been removed, we will confirm later that the reconstructed speech have a better intelligibility than the noisy speech.

4.4. Performance Evaluation

In order to show the effectiveness of the proposed algorithm, we compare it with scNMF and statistic method which is very powerful for normal speech enhancement. Specifically, power subtraction method proposed in [18] and optimal gain modification based method proposed in [19] were both used to enhance the noisy speech for further comparison.

Figure 4 shows the spectrums of the enhanced speech using different algorithms. Figure 4 (a) plots the spectrum of the clean speech. Figure 4 (b) plots the spectrum of the noisy speech corrupted by F16 fighter jet noise at SNR of -5 dB. Figure 4(c) plots the spectrum of the estimated clean speech enhanced using the proposed algorithm. Figure 4 (d) plots the spectrum of the estimated clean speech enhanced using the algorithm proposed by Smaragdis. Figure 4(e) plots the spectrum of the estimated clean speech enhanced by the power subtraction, and Figure 4(f) plots the spectrum of the estimated clean speech enhanced by the statistic like algorithm proposed by Cohen [19]. From Figure 4, we can find that, compared with scNMF and power subtraction algorithm, the spectrum of the estimated clean speech enhanced by the proposed algorithm retains more speech content. Compared with the statistic algorithm proposed by Cohen, more noise has been subtracted and more spectrum components have been retained using the proposed algorithm.

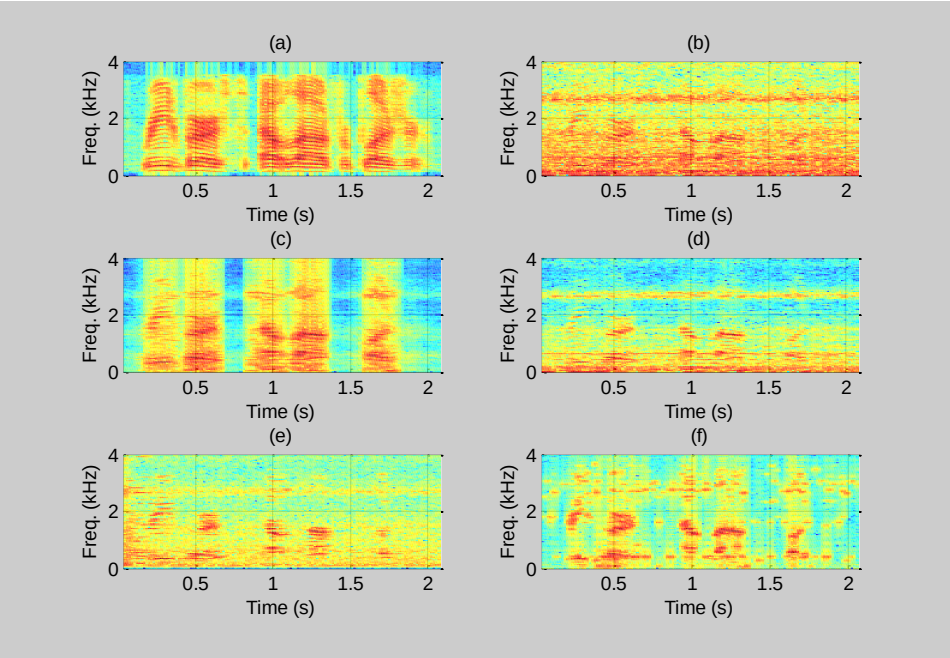

Figure 4. Spectrums of Enhanced Speech using Different Algorithms. (a) Clean Speech. (b) Noisy Speech Corrupted by F16 Fighter Jet Noise at SNR of -5 dB. (c) Speech Enhanced using the Proposed Algorithm. (d) Speech Enhanced using the Algorithm Proposed by Smaragdis. (e) Speech Enhanced by the Algorithm Proposed by Cohen [19], and (f) Speech Enhanced by the Power Subtraction

Ten listeners were recruited for the listening tests with half of each gender. In the identification test, stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. The three processing conditions included the noise corrupted speech (denoted as UN), noisy speech processed using the proposed algorithm (denoted as pscNMF), synthesized speech using the conventional sparseness constraint convolutive NMF algorithm (denoted as scNMF), synthesized speech using the power subtraction algorithm (denoted as PS). In addition, the algorithm proposed by Cohen in [19] (denoted as OMLSA) was also evaluated for its ability to improve speech intelligibility. The OMLSA algorithm minimizes the mean-square error of the log-spectra based on a Gaussian statistical model associated with the speech presence uncertainty. The noise spectrum is estimated by recursively averaging past spectral power values, using a smoothing parameter that is adjusted by the speech presence probability in subbands. In this paper, the parameters of OMLSA were the same as that in [19].

Table 1. Word Identification Rates of Different Stimuli with Different SNRs

		Performance (%)				
SNR(dB)	Noise	UN	pscNMF	cNMF	OMLSA	PS
-10	GWN	53.82	58.74	57.77	49.50	47.38
	Babble	57.87	64.57	56.83	49.23	44.01
	F16	51.23	59.22	50.51	48.38	40.51
	GWN	60.82	69.33	65.81	49.31	59.09

-5	Babble	65.10	70.05	64.35	57.80	54.91
	F16	59.26	68.08	59.35	50.27	59.29
	GWN	68.86	75.59	75.36	62.09	71.82
0	Babble	72.85	73.12	71.92	68.95	66.44
	F16	69.25	75.32	67.95	66.12	71.51

The duration of each sentence was approximately two seconds. The experiments were performed in a sound-proof room. Prior to the sentence test, each subject listened to a set of noise-masked sentences to become familiar with the testing procedure. Five-minute breaks were given to the subjects every 30 minutes. A total of 30 sentences were used per condition in each trial (a total of ten trials). The order of the conditions was randomized across subjects. Listeners were asked to write down the words they heard, and intelligibility performance was assessed by counting the number of words identified correctly.

Table 1 shows the word identification rates of unprocessed noisy speech and its enhanced version using different algorithms in different noise contexts with different SNR levels. As can be seen from Table 1, the recognition rate of the estimated clean speech using pscNMF has substantial high identification rate than the unprocessed noisy speech and that obtained using other speech enhancement algorithms in different SNR levels.

5. Conclusion

Improving speech intelligibility is a key issue when conducting speech enhancement in low SNR environment. Conventional speech enhancement algorithms fail to tackle this problem. In this paper, we proposed a sparseness constraint based non-negative matrix factorization algorithm and applied it to conduct speech intelligibility enhancement. Experimental results show that the intelligibility of the speech enhanced using the proposed algorithm was substantially higher than that of the unprocessed noisy speech and that of the conventional speech enhancement algorithms. Extensive comparisons demonstrate that the system has gained the state-of-the-art performance in speech intelligibility enhancement.

Acknowledgements

This work is partly supported by the Natural Science Foundation of China (61301295, 61301219, 61003131), the Natural Science Foundation of Anhui Province (1308085QF100, 1408085MF113) and Doctoral Fund of Anhui University.

References

[1] P.C. Loizou, “Speech enhancement: theory and practice”, CRC, New York **(2007).**
[2] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners”, The Journal of the Acoustical Society of America, vol. 126, no. 3, **(2009).**
[3] A. S. Bregman, “Auditory scene analysis: The perceptual organization of sound”, The MIT Press, **(1994).**
[4] M. Cooke and D. P. W. Ellis, “The auditory organization of speech and other sources in listeners and computational models”, Speech Communication, vol. 35, pp. 3-4, (**2001**).
[5] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction”, The Journal of the Acoustical Society of America, vol. 123, no. 3, **(2008).**
[6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization”, Nature, 401, 6755, **(1999).**

[7] K. J. Hayworth and I. Biederman, "Neural evidence for intermediate representations in object recognition", Vision research, vol. 46, no. 23, **(2006).**
[8] D. L. Wang and G. J. Brown, "Computational auditory scene analysis: Principles, algorithms, and applications", IEEE Press, **(2006).**
[9] R. Weiss and J. Bello, "Unsupervised discovery of temporal structure in music", IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 6, **(2011).**
[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 3, **(2007).**
[11] P. OGrady and B. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity", Independent Component Analysis and Signal Separation, **(2007).**
[12] Y. C. Cho, S. Choi and S. Y. Bang, "Non-negative component parts of sound for classification", Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology, **(2003)**, December 14-17.
[13] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, **(2007).**
[14] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization", Advances in neural information processing systems, vol. 13, **(2001).**
[15] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs", Independent Component Analysis and Blind Signal Separation, **(2004).**
[16] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek and M. Weinstock, "IEEE recommended practice for speech quality measurements", IEEE Trans. Audio Electroacoust, vol. 17, no. 3, **(1969).**
[17] 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, **(1993).**
[18] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, **(1979).**
[19] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments", Signal processing, vol. 81, no. 11, **(2001).**

Authors

Jian Zhou, He received his B. S. degree from Southwest Jiaotong University, China (2004) and M. S. degree from Southwest Jiaotong University, China (2007). He received the Ph.D. degree from Southeast University, China, in 2013. His research interests include speech enhancement, whispered speech reconstruction, blind source separation, image processing and pattern recognition.

Xianyong Fang, He received M. S. degree from Hefei University of technology, China (2007). He received the Ph.D. degree from state Key Laboratory of CAD and CG, Zhejiang University, China, in 2005. Currently, he is a professor at the school of computer science and technology, Anhui University, China. His research interests include computer graphics and pattern recognition.

Liang Tao, He received M. S. degree from Anhui University, China (1998). He received the Ph.D. degree from University of Science and Technology of China in 2005. Currently, he is a professor at the school of computer science and technology, Anhui University, China. His research interests include signal processing and pattern recognition.

Li Zhao, He received his B. S. degree from Nanjing University of Aeronautics and Astronautics, China (1982) and M. S. degree from Southeast University, China (1988). He received the Ph.D. degree from Kyoto Institute of Technology, Japan, in 1995. Currently, he is a professor at the College of information science and engineering, Southeast University, China. His research interests include speech processing, audio and video signal processing.