**International Journal of Electrical and Electronics Research** ISSN 2348-6988 (online)
Vol. 8, Issue 1, pp: (9-15), Month: January - March 2020, Available at: www.researchpublish.com

# Non-stationary additive noise signal filtration process in NMF based approach for the single-channel speech enhancement

[1]Ravi Shankar Prasad, [2]Pradeep singh yadav

M. Tech Scholar, Assistant professor (Electronic and Telecommunication) SSITM

Ravishankar.parased@gmail.com, pradeepyadav.py3@gmail.com

***Abstract:*** **This paper investigates a non-negative matrix factorization (NMF)-based approach to the semi-supervised single-channel speech enhancement problem where only non-stationary additive noise signals are given. The NMF spectral basis matrices for both speech and noise are obtained in a manner of supervised learning, and thus the performance of their associated NMF speech enhancement degrades as the speaker and/or noise characteristics are not matched for the learning and evaluation environment. The experimental evaluation was made on TIMIT corpus mixed with various types of noise. It has been shown that the proposed method outperforms some of the state-of-the-art noise suppression techniques in terms of signal-to-noise ratio.**

***Keywords:*** **non-negative matrix factorization (NMF), noise suppression techniques, signal-to-noise ratio.**

## 1. INTRODUCTION

Signal means information and processing means operation. It means how information in the form of signal is operated or modified to get desired signal and how system process these signal[1].

Signal processing is very wide field. We are all immersed in a sea of signals. All of us from the smallest living unit, a cell, to the most complex living organism (humans) are all time receiving signals and are processing them. Survival of any living organism depends upon processing the signals appropriately. What is signal? To define this precisely is a difficult task. Anything which carries information is a signal. In this course we will learn some of the mathematical representations of the signals, which has been found very useful in making information processing systems[2]. Examples of signals are human voice, chirping of birds, smoke signals, gestures (sign language), and fragrances of the flowers. Many of our body functions are regulated by chemical signals, blind people use sense of touch. Bees communicate by their dancing pattern.

Speech enhancement in presence of background noise is an important problem that exists for a long time and still is widely studied nowadays. The efficient single-channel noise suppression (or noise reduction) techniques are essential for increasing quality and intelligibility of speech, as well as improving noise robustness for automatic speech recognition (ASR) systems[3]**.** Generally speaking, the aforementioned techniques can be a form of machine learning, which is usually divided into two categories: supervised learning and unsupervised learning[4]. In supervised learning the training data are well classified and labeled while in unsupervised learning they are not.

By processing we mean operating in some fashion on a signal to extract some useful information. For example when we hear same thing we use our ears and auditory path ways in the brain to extract the information[3]. The signal is processed by a system. In the example mentioned above the system is biological in nature. We can use an electronic system to try to mimic this behavior. The signal processor may be an electronic system, a mechanical system or even it might be a computer program. The word digital in digital signal processing means that the processing is done either by a digital hardware or by a digital computer.

Speech enhancement or noise reduction has been one of the main investigated problems in the speech community for a long time. The problem arises whenever a desired speech signal is degraded by some disturbing noise[5].Speech enhancement algorithms can be also used to design robust speech/speaker recognition systems by reducing the mismatch between the training and testing stages. In this case, a speech enhancement approach is applied to reduce the noise before extracting a set of features

In this paper, we address another issue regarding the NMF-based speech enhancement, which is the mismatch between the learning and evaluation situations. Speaking in more detail, the traits of speech and/or noise in the NMF learning processing are not necessarily close to those in the utterances to be enhanced[6]. Such a mismatch usually leads to performance drop of the NMF enhancement. Experimental results showed that the proposed algorithm outperformed not only the statistical model based and NMF-based methods but also the combination of them.

## 2. RELATED WORK

This paper describes a method for enhancing speech corrupted by broadband noise. The method is based on the spectral noise subtraction method. The original method entails subtracting an estimate of the noise power spectrum from the speech power spectrum, setting negative differences to zero, recombining the new power spectrum with the original phase, and then reconstructing the time waveform.

To conclude, the main differences between the basic spectral subtraction method and our implementation are that we subtract an overestimate of the noise spectrum and prevent the resultant spectral components from going below a central floor. We consider a sensor array located in an enclosure, where arbitrary transfer functions (TFs) relate the source signal and the sensors. The array is used for enhancing a signal contaminated by interference. Constrained minimum power adaptive beam forming, which has been suggested by Frost and, in particular, the generalized side lobe canceller (GSC) version, which has been developed by Griffiths and Jim, are the most widely used beam forming techniques.

The suggested algorithm can be applied for enhancing an arbitrary nonstationary signal corrupted by stationary noise. An arbitrary TF and array geometry can be used. The use of TFs ratio rather than the TFs themselves (which is the counterpart of relative delay in delay-only arrays) improves the efficiency and robustness of the algorithm since shorter filters can be used. Although our algorithm was implemented in the frequency domain, it can also be implemented in the time domain. This applies both to the adaptive beam former stage and to the system identification stage.

In this paper, we have assumed that the noise is nonstationary. Sometimes, this assumption is not accurate (e.g., for a cocktail party noise). Nevertheless, whenever the noise is “more stationary” compared with the desired speech signal, the estimation method presented in Section IV is expected to be useful. In order to use the proposed algorithm, one needs to re-estimate the TFs once the acoustic environment has changed. In order to reduce the computational complexity, recursive procedures [e.g., RLS methods for solving (30)] may be incorporated. This is left as a further research topic.

We propose a new speech enhancement method based on the time adaption of wavelet thresholds. The time dependence is introduced by approximating the Teaser Energy of the wavelets coefficients. To our knowledge, the proposed method is one of the first successful applications of the wavelet thresholding method for speech enhancement.

We report on the development of a noisy speech corpus suitable for evaluation of speech enhancement algorithms. This corpus is used for the subjective evaluation of 13 speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener algorithms. Of the two subspace algorithms examined, the generalized subspace approach [8] performed consistently better in OVRL scale across all SNR conditions and four types of noise.

We propose, an improved form of iterative speech enhancement for single channel inputs is formulated. The basis of the procedure is sequential maximum a posteriori estimation of the speech waveform and its all-pole parameters as originally formulated by Lim and Oppenheim, followed by imposition of constraints upon the sequence of speech spectra. The problem of enhancing speech degraded by additive white and slowly varying colored background noise was addressed. In addition, algorithm performance as a preprocessor for speech recognition was also considered.

We previously have applied deep auto encoder (DAE) for noise reduction and speech enhancement. However, the DAE was trained using only clean speech. In this study, by using noisy clean training pairs, we further introduce a demising process in learning the DAE. In training the DAE, we still adopt greedy layer-wised retraining plus fine tuning strategy. Deep learning has been successfully applied in pattern classification and signal processing, particularly in acoustic modeling for ASR. Based on the same idea, we have applied the DAE for noise reduction and speech enhancement [7]. In this study, we further introduced a demising processing in training the AE by using noisy-clean speech pairs.

Many issues need to be further investigated. The first one is how to effectively incorporate prior knowledge in modeling the DAE. For example, speech signal has many well-structured, multi-scale temporal-frequency patterns and transitions. Reducing the interference noise in a monaural noisy speech signal has been a challenging task for many years. Compared to traditional unsupervised speech enhancement methods, e.g., Wiener filtering, supervised approaches, such as algorithms based on hidden Markov models (HMM), lead to higher-quality enhanced speech signals.

This paper investigated the application of NMF in speech enhancement systems. We developed speech enhancement methods using a Bayesian formulation of NMF (BNMF). We proposed two BNMF-based systems to enhance the noisy signal in which the noise type is not known a priori. We present a technique for denoising speech using nonnegative matrix factorization (NMF) in combination with statistical speech and noise models. We compare our new technique to standard NMF and to a state-of-the-art Wiener filter implementation and show improvements in speech quality across a range of interfering noise types.

We have shown that NMF can be used to denoise speech in the presence of non-stationary noise, and we have shown that by regularizing NMF based on a prior model of speech and noise, we can exploit additional signal structure to improve performance. There are a number of interesting directions for future work. This work complements work on explicit control of sparseness for source-separation [9], and combining the two approaches may improve results further.

This paper proposes Hidden Markov Model (HMM) for speech enhancement. The proposed model is based upon the combination of the hidden Markova model (HMM) with the non-negative matrix factorization (NMF).The proposed model has been designed for the speech signal enhancement using the combination of hidden Markova model with the non-negative matrix factorization (HMM-NMF). The supervised non-negative matrix factorization model has been used for the sparse matrix formation.

We propose a dual-microphone speech enhancement framework based on β-NMF in the paper. This method extends single-microphone speech enhancement based on NMF by introducing the interchange matrix to the cost function. In this paper, we propose to train a classifier in order to overcome such poor characterization of the signals through the trained models. The main idea is to decompose the noisy observation into parts and the enhanced signal is reconstructed by combining the less-corrupted ones which are identified in the kestrel domain using the trained classifier.

Single-channel speech enhancement method for suppressing non-stationary noise in a noisy speech signal has been presented. We have tried to overcome the problem of missing training data by using a trained classifier instead of trained models based on NMF .In this paper, we introduce a training and compensation algorithm of the class-conditioned basis vectors in the non-negative matrix factorization (NMF) model for single-channel speech enhancement.

We introduced a training and compensation algorithm of the class-conditioned basis vectors in the NMF model for single channel speech enhancement. We considered the PGM for both the NMF and classification models. Finally, we comment on some interesting research avenues for further improving the enhancement performance of our proposed method. Firstly, we can consider modeling the basis vectors using a more accurate multimodal distribution

This letter presents a speech enhancement technique combining statistical models and non-negative matrix factorization (NMF) with on-line update of speech and noise bases. The statistical model-based enhancement methods have been known to be less effective to non-stationary noises while the template-based enhancement techniques can deal with them quite well. This letter has proposed a speech enhancement technique combining statistical model-based and NMF approaches with on-line update of speech and noise bases. The combination of two distinct approaches in conjunction with the on-line bases update enables an efficient suppression of non-stationary noises even with mismatched training data.

This paper presents a statistical method of single-channel speech enhancement that uses a variation auto encoder (VAE) as a prior distribution on clean speech. A standard approach to speech enhancement is to train a deep neural network (DNN) to take noisy speech as input and output clean speech. We presented a semi-supervised speech enhancement method, called VAE-NMF that involves a probabilistic generative model of speech based on a VAE and that of noise based on NMF. Only the speech model is trained in advance by using a sufficient amount of clean speech. One interesting future direction is to extend VAE-NMF to the multichannel scenario. Since complicated speech signals and a spatial mixing process can be represented by a VAE and a well-studied phase-aware linear model

## 3. METHODOLOGY

This section considers speech $D_s$and noise $D_n$matrix dictionaries in which columns or atoms follow specific linear models that are discussed below.

### 3.1. Speech model

The basic of speech production assumes that the time-domain excitation source $e(t)$and vocal tract filter $a(t)$ are combined into convoluted model

$$s(t) = a(t) * e(t)$$

The excitation signal e(t) itself could be presented by summing up complex sinusoids and noise on frequencies that are multiples of fundamental frequency

$$e(t) = \sum_{k=1}^{p} c_k \, exp\big(ik\bar{w}(t)\big)$$

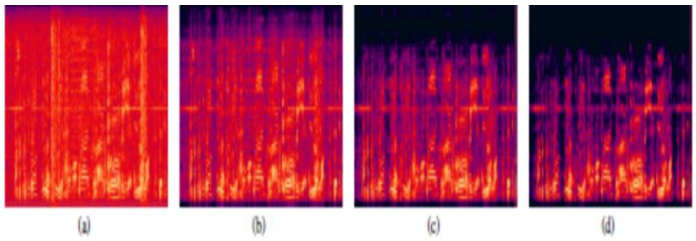

**Figure 1: Spectrograms of (a) noisy signal, and denoised variant using (b) linNMF, (c) denseNMF and (d) denseNMF with higher sparsity constraints**

The quasi-stationarity of the transfer function $a(t$ )and fundamental frequency function $\bar{w}(t$ )takes place during the short frame$t \in T_T$, so that $a(t) = a_T(t)$and $\bar{w}(t) = \bar{w}_r$. Putting together two representations (1) and (2) provides:

$$s_r(t) = \sum_{k=1}^{p} ck\hat{a}_r \; (k\bar{w}_r) exp\big(ik\bar{w}_r(t)\big)$$

where the hat symbol indicates Fourier transform of the corresponding function. The approximation of signal (3) in magnitude Short-Time Fourier Transform (STFT) domain using window function w(t) is given as:

$$|\hat{s}_t(w)| \approx \sum_{k=1}^{p} ck|\hat{a}_r(k\bar{w}_r)||\hat{w}(w - k\bar{w}_r)|$$

Since $\hat{w}(w)$localizes most energy at low frequencies. Then we suppose that frequency response $|\hat{a}(w)|$at any frame τ is modeled by the (possibly sparse) combination of fixed spectral shapes with non-negative weights $x_j^s(T) \geq 0, i.e. |\hat{a}_t(w)| = \sum_{j=1}^{m} x_j^s\,(T)\big|\hat{a}_j(w)\big|$ |. In this case the final representation formula is:

$$|s\hat{}_r\,(w)| \approx \sum_{j=1}^{m} x_j^s(r) \sum_{k=}^{p} c_k\big|\hat{a}_j(k\bar{w}r)\big||\hat{w}(w - k\bar{w}r)|$$

The latest representation could be efficiently wrapped in matrix notation by taking discrete values $\{w_i\}_{i=1}^{k}$ and$\{T_l\}_{l=1}^{T}$, since only one fundamental frequency $\bar{w}_r$ is expected for the given time frame. It is possible to choose it from the discrete set $\bar{\cap}_L$of L hypothesized fundamental frequencies bounded by $\bar{w}_{min}$ and$\bar{w}_{max}$. This leads to $m_{s=Lm}$ possible combinations. Therefore the following equation for the input speech spectrogram $y_s \in R_+^{K\times T}$ holds:

$$\begin{cases} y_s = D_s X_s \\ d_j^s = W_j C a_j \end{cases} \quad j = 1,2 \ldots \ldots m_s$$

Each column $d_j^s$ of matrix $D_s \ \epsilon R_+^{K\times m_s}$ represents one harmonic atom, in which isolated harmonics are placed in columns of matrices $w_j \epsilon R_+^{K\times P}$, weighted by constant amplitude matrix $C = diag(c_1 \ldots\ldots c_p)$. The representation coefficients $a_j \epsilon R_+^{K\times P}$ and gain matrix $X_s$ are needed to be defined.

### 3.2. Noise model

Being not so physically motivated as speech model the noise model of signal is constructed in the similar way by assuming additives of corresponding spectral shapes (however this could also be theoretically approved by exploring band-limited noise signals, that is not the case of the current work). Each time varying noise magnitude $|\hat{n}_r(w)|$ is decomposed into sum of individual static components:

$$|\hat{n}_r(w)| = \sum_{k=1}^{r} b_{k(r)} |\hat{n}_k(w)|$$

Note that $|\hat{n}_k(w)|$ is the predefined set of noise magnitudes extracted from the known noise signals, whereas non-negative filter gains $b_{k(r)} \geq 0$ should be defined from the observed data. As stated before the non-negative combination $b_k^n(r) = \sum_{j=1}^{m_n} x_j^n(r) b_{kj}$ leads to the following representation:

$$|\hat{n}_r(w)| = \sum_{j=1}^{m_n} x_j^n(r) \sum_{k=1}^{r} |\hat{n}_k(w)| b_{kj}$$

And in matrix notation for the same discrete $w_i$ and $r_l$ on input noise spectrogram $y_n \epsilon R_+^{K\times T}$

$$\begin{cases} y_n = D_n X_n \\ d_j^n = N b_j, \end{cases} j = 1,2 \ldots\ldots m_n$$

With $D_n \ \epsilon R_+^{K\times m_n}$ represents noise dictionary with atoms $d_j^n, N \ \epsilon R_+^{K\times r}$ contains noise spectral shapes combined with unknown coefficients $b_j \epsilon R_+^r$ to produce noise model.

### 3.3. NMF with linear constraints

Here we introduce the general formulation of NMF problem with linear constraints that follows from speech (6) and noise (9) representations. As soon as spectrograms in both cases factorize by the product of two non-negative matrices, the β-divergence between observed spectrogram Y and $DX$ product could be chosen for approximation [13]. Here we only consider a special case called Kullback-Leibler divergence $D_{KL}(Y\|DX) = \sum_{i,j} Y_{i,j} \log \frac{Y_{ij}}{(DX)_{ij}} - Y_{ij} + (DX)_{ij}$, but other divergences could be used as well.

The following general optimization problem $(linNME)$ gives solution to (6) and (9)

$$\begin{cases} D_{KL}(Y\|DX) + \lambda\|X\|1 \to min \\ d_{ij=\Psi_j a_j,} \quad j = 1,2, \ldots\ldots m \end{cases}$$

with minimization over $\{a_j\}_{j=1}^m$ and $X$. The factors dj are spanned by the columns of matrices Ψj that could be diverse for j = 1, 2 . . . m. Using multiplicative updates heuristic it could be shown that the following algorithm solves the optimization problem.

$$a_j \longleftarrow a_j \cdot \frac{\Psi_j^T \frac{Y}{DX} \overline{x}_j^T}{\Psi_j^T \mathbf{1} \overline{x}_j^T}$$

$$X \longleftarrow X \cdot \frac{D^T \frac{Y}{DX}}{(D^T \mathbf{1} + \lambda)}$$

where $\bar{x}_j$denotes $j = th$ row of matrix$X$, $1\ \epsilon\ R^{K\times T}$is the all-ones matrix and multiplications $a.b$ and divisions $\frac{x}{y}$ are element-wise. By iterating these rules the factors dj are modeled in linear subspace with dimensionality implied by rank of $\Psi_j$ .

### 3.4. NMF with linear dense constraints

It has been experimentally found that many solutions achieved by (11) tend to "reduce" the rank of corresponding subspace. In other words, each dj tends to have sparse representation in basis $\Psi_j$ . It is not the desired solution in the current task, in case if $\Psi_j$ contains windowed-sinusoid magnitude values on the particular frequency. As we want to extract full harmonic atoms from signal, it is expected that every harmonic has non-zero amplitude (especially in low-frequency band). The denseNMF optimization task is proposed to overcome this problem that favors non-zero coefficients in vectors$\{a_j\}_{j=1}^{m}$:

$$\begin{cases} D_{KL}(Y\|DX) + \lambda\|X\|_1 + a\sum_j \|a_j\|_2^2 \rightarrow min \\ d_{ij=}\Psi_j a_{j,}\|a_j\|_1 = 1 \qquad j = 1,2 \ldots\ldots m \end{cases}$$

$$\tilde{a}_j = a_j/\|a_j\|_1:$$

The following rules are also derived from multiplicative updates for l1-normalized coefficients

$$a_j \leftarrow \bar{a}_j \frac{1_j \tilde{a}_j^T \Psi_j^T 1\bar{x}_j^T + a1_j \bar{a}_j^T \bar{a}_j}{\Psi_j^T 1\bar{x}_j^T + 1_j \tilde{a}_j^T \Psi_j^T \frac{Y}{DX}\bar{x}_j^T + a\tilde{a}_j}$$

$$X \leftarrow X . \frac{D^T \frac{Y}{DX}}{(D^T 1 + \lambda)}$$

Where $1_j$ indicates the vector of all-ones of the same size as$a_j$. It should be noted that convergence properties of presented algorithms (11) (13) are not studied. However during the experiments the monotonic behavior of the target function (10) has been permanently observed for arbitrary $a > 0$.

## 4. RESULTS

Selecting the clean conversation and the sound is the TIMIT database [25] and the NOISEX data source [26], where using down sampling we can change the sampling rate of most signals to 8 kHz. With this study, working out for the clean talk consists of 20 sentences (60 mere seconds) pronounced by 10 men and 10 females. Each one of the test speech indicators for the speech improvement work is one phrase. We choose two background noises in the paper: the Hf route and Factory1 sounds. Besides, training data and test data in the test are disjoint. For the proposed framework, the windowpane function, the applied body size, and the frame change are Hamming home window, 512 samples and 128 samples, respectively. Based on the standard decision of $K \leq IJ/(I + J)$, presuming the clean conversation and noise basis vectors, K is defined to 30, respectively, and allow maximum iteration quantity be add up to 50. Both microphones with a 4 cm spacing distance found noisy speech signals which were produced by convolving the prospective and noise resources with a couple of HRTFs assessed in the mildly reverberant room (T60 ≈ 220ms) with sizes 4.3 × 3.8 × 2.3 m3 (size ×width × elevation), with the addition of the sound to the clean screening speech to create the noisy indicators at four transmission to-noise ratios (SNRs): −10, −5, 0, and 5 dB. The length between the focus on source and the midpoint of both microphones is set to at least one 1.2 m. The path of introduction (DOA) was chosen, respectively, relating to to $\theta \propto \{0^0, 30^0, 60^0, 90^0, 120^0, 150^0, 180^0\}$ .The squared Euclidean distance β = 2 is utilized for simplicity.

## 5. CONCLUSION

The primary contributions of the paper are pursuing. First we have applied prior deterministic modeling of talk and noise signals inside NMF-based speech enhancement framework. They have resulted in linearly constrained columns or atoms of related dictionaries Ds and Dn. Then we've suggested the new optimization problem declaration and multiplicative improvements algorithm that regularizes the representation coefficients in order that they contain as fewer zeros as you possibly can, i.e., acquiring thick solution. We've examined the new method on TIMIT corpus blended with sounds on different SNR, attaining the best result for low SNR among some state-of-the-art sound suppression algorithms, and

somewhat outperforming ”oracle” NMF estimator with known clean indicators. In the future, the proposed model can be enhanced using the combination of the other effective techniques along with the proposed model such as linear component analysis or log-spectral amplitude for the improvement in the signal enhancement model. Single-microphone speech enhancement algorithms by using nonnegative matrix factorization can only utilize the temporal and spectral diversity of the received signal, making the performance of the noise suppression degrade rapidly in a complex environment.

## REFERENCES

[1] K. Kwon, J. W. Shin, and N. S. Kim, “NMF-Based Speech Enhancement,” IEEE Signal Process. Lett., vol. 22, no. 4, pp. 450–454, 2015.

[2] P. Sinha, “Speech processing in embedded systems,” Speech Process. Embed. Syst., pp. 1–171, 2010.

[3] M. Bahoura and J. Rouat, “Wavelet speech enhancement based on the Teager energy operator,” IEEE Signal Process. Lett., vol. 8, no. 1, pp. 10–12, 2001.

[4] E. Alpaydın, Introduction to machine learning. 2014.

[5] M. Berouti, R. Schwartz, and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” no. May 1979, pp. 208–211, 2005.

[6] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” vol. 5, pp. 1457–1469, 2004.