Non-negative Tensor Factorization for Speech Enhancement

Liang He[1,*], Weiqiang Zhang[1] and Mengnan Shi[1]
[1]Department of Electronic Engineering, Tsinghua University, China
*Corresponding author

***Abstract*-This paper proposes an algorithm for speech enhancement by non-negative tensor factorisation. We group adjacent time-frequency matrices in the spectrograms together to form a tensor as a basic input in our algorithm. The non-negative tensor factorisation is followed to perform sound source separation between speeches and noises. The proposed strategy benefits from both short time spectral analysis and long term information. From the consideration of auditory theory and linguistics, the latter preserves the temporal dynamics information and intrinsic structure of speech, which are important for the continuity and integrity of hearing. We collected several types of real-life noises and conducted experiments on the TIMIT database. Experimental results demonstrated that the segmental signal to noise ratio (SSNR) and the perceptual evaluation of speech quality (PESQ) were significantly improved respectively.**

Keywords-non-negative tensor factorization (NTF); speech enhancement; sound source separation

I. INTRODUCTION

Speech enhancement is crucial for real-life applications and still a challenging task today. Most traditional speech enhancement algorithms base on short time spectral amplitude and operate in a frame-by-frame way [1][2]. Although these algorithms have good performance and are widely used in many applications, they fail to benefit from long-term information. The long-term information often exists in phonetic, prosodic, lexical and *etc*. features which are essential for auditory perception [3]. However, as far as we know, speech enhancement algorithms based on long term features often have worse performance compared with the ones based on spectral features. In order to solve this dilemma, we plan to bring more long-term information in our spectral feature based speech enhancement algorithm.

Recently, deep neural network (DNN) achieved great success in the field of speech recognition [4][5]. Although the fundamental mathematic and cognition explanation for this breakthrough is still in the discussion, there are several recognized causes. We believe one of them is the long-term spectral features for statistical modeling. From the consideration of language's phonology, phoneme is a basic unit and the smallest contrastive linguistic unit. The time duration of a typical phoneme is tens to hundreds of milliseconds. For a spectral feature based algorithm, we often set the frame duration 10-30 milliseconds to obtain short time stationary property [3]. So, a single spectral frame can't cover a phoneme in most cases, which means a single frame lacks necessary information. Considering other auditory factors such as the auditory masking effect, an input with too short duration is improper and long term spectral features are needed. A simple way to make the long term feature is by stacking several adjacent spectral feature vectors to form a large vector. The existing algorithms can take the stacked vectors as inputs without any modification. However, this strategy violates intrinsic structure of speech and benefits from the long term information in a low level. We group several adjacent time-frequency matrices in the spectrogram together to form a tensor as our basic unit for statistical modeling. It incorporates the benefits both from the short time Fourier analysis and long term information in a novel way.

Since the feature representation is a tensor rather than a vector or matrix, the statistical modeling should be adjusted accordingly. Non-negative tensor factorization (NTF) is a natural extension of non-negative matrix factorization (NMF) [6]. NMF factorize a matrix V into two matrix W and U with the property that all three matrices have no negative elements. In the field of image processing, speech signal processing, text analysis and *etc*, NMF is more inherent to the data compared with other related algorithms, such as principle component analysis (PCA).

Wilson and *et.al.* first proposed a NMF based speech enhancement algorithm and obtained good results[7]. Mohammadiha and *et.al.* put forward this idea in several ways, for example the design of filter model, the study of NMF constraint condition and the comparison of supervised and unsupervised methods, and further proved the effectiveness of the NMF in the field of speech enhancement [8][9][10]. In spite of these improvement, they still take vectors as inputs. Due to the above analysis, the tensors are more suitable for the speech enhancement. So our mathematical tool is changed from the NMF to the NTF. The NTF is capable of processing tensors and preserves good properties of the NMF. Hence it promises a better performance. Fitzgerald brought the NTF to sound source separation [11]. Barker performed separation by a Wiener-like filter generated from the estimated tensor factors and obtained better results [12]. Gemmeke proposed exemplar-based sparse representations for speech recognition in a noisy condition. The underlying idea is similar to our method [13].

The remainder is as follows: Section II presents NTF, section III proposes our NTF based speech enhancement algorithm, and section IV describes our experiments on the

TIMIT database and analyzes results. Finally, a conclusion is summarized in section V.

II. Non-negative Tensor Analysis

We adopt the PARAFAC (Parallel Factor Analysis) model to represent a tensor for its simplicity [6]. The PARAFAC model decomposes a tensor into a sum of multi-linear terms. Let a tensor $\mathcal{X}$ be a three-way array ($R \times N \times M$). According to the PARAFAC model, the $\mathcal{X}$ consists of a product of three components: $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_K]$, $B = [\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_K]$, $C = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_K]$ and a residual tensor $\mathcal{E}$, see Figure 1

$$\mathcal{X} = \sum_{k=1}^{K} \boldsymbol{a}_{:k} \circ \boldsymbol{b}_{:k} \circ \boldsymbol{c}_{:k} + \mathcal{E}$$

where K is the parallel component number, $\circ$ is the outer product and $(\cdot)_{:k}$ denote the k-th column of matrix $(\cdot)$. We choose the minimization of the standard squared Euclidean distance (Frobenius norm) as our objective function to estimate the non-negative elements in components A, B and C

$$\begin{aligned} & \min \|\mathcal{E}\|_F^2 \\ \text{s.t.} \quad & a_{rk} \geq 0, b_{nk} \geq 0, c_{mk} \geq 0, \forall r, n, k, m \\ & \|\boldsymbol{a}_{:k}\| = 1, \|\boldsymbol{b}_{:k}\| = 1, \forall k \end{aligned}$$

To solve the above optimization problem, there are at least three different approaches. The first one uses vectorization to form a large vector and employs traditional methods to solve it. This idea is straight but requires extremely high memory and computation. The second one uses non-linear conjugate gradient optimization. However, for the non-convexity of objective function, this approach fails to reach an optimal solution in most cases. The most popular approach is an alternating least squares (ALS). The core idea of the ALS is to compute the gradient of objective function with respect to each component matrix and update each component matrix in an alternating and iterative way. Although the ALS has many advantages, taking some weak conditions (over-determined case or under-determined case) and computation issue into account, we adopt the hierarchical alternating least squares (HALS) algorithm to handle the optimization problem. The HALS break the objective function into a set of simple functions and minimize them sequentially.

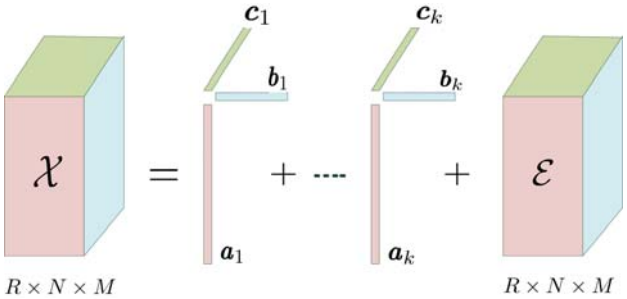

FIGURE I . PARAFAC MODEL

Consider sequential minimization of the local cost functions $f^{(k)}$

$$f^{(k)} = \|\mathcal{X}^{(k)} - \boldsymbol{a}_{:k} \circ \boldsymbol{b}_{:k} \circ \boldsymbol{c}_{:k}\|_F^2$$

where $\mathcal{X}^{(k)} = \mathcal{X} - \sum_{j=1, j\neq k}^{K} \boldsymbol{a}_{:j} \circ \boldsymbol{b}_{:j} \circ \boldsymbol{c}_{:j}$. The gradient of $f^{(k)}$ with respect to element $\boldsymbol{a}_{:k}$ is given by

$$\frac{\partial f^{(k)}}{\partial \boldsymbol{a}_{:k}} = -\mathcal{X}_a^{(k)} \boldsymbol{b}_{:k} \odot \boldsymbol{c}_{:k} + \boldsymbol{a}_{:k} (\boldsymbol{b}_{:k} \odot \boldsymbol{c}_{:k})^t (\boldsymbol{b}_{:k} \odot \boldsymbol{c}_{:k})$$

where $\odot$ is the Khatri-Rao product and $\mathcal{X}_a$ is mode- a matrixized version of $\mathcal{X}$ by unfolding. In a similar way, we can compute $\frac{\partial f^{(k)}}{\partial \boldsymbol{b}_{:k}}$, $\frac{\partial f^{(k)}}{\partial \boldsymbol{c}_{:k}}$ and reach an iterative optimization algorithm by setting the gradients to zero, see Algorithm 1. More details can be found in the reference [6].

Algorithm 1 HALS NTF

Input: $\mathcal{X}$ and K
Output: non-negative matrix A, B and C such that the objective function is minimized

random non-negative initialization for A, B and C
for $k = 1$ to K **do**
  $\boldsymbol{a}_{:k} = \boldsymbol{a}_{:k} / \|\boldsymbol{a}_{:k}\|^2$, $\boldsymbol{b}_{:k} = \boldsymbol{b}_{:k} / \|\boldsymbol{b}_{:k}\|^2$
end for
$\mathcal{E} = \mathcal{X} - \sum_{k=1}^{K} \boldsymbol{a}_{:k} \circ \boldsymbol{b}_{:k} \circ \boldsymbol{c}_{:k}$
while a stopping criterion is not met **do**
  for $k = 1$ to K **do**
    $\mathcal{X}^{(k)} = \mathcal{E} + \boldsymbol{a}_{:k} \circ \boldsymbol{b}_{:k} \circ \boldsymbol{c}_{:k}$
    $\boldsymbol{a}_{:k} \leftarrow \left[\mathcal{X}_a^{(k)} \boldsymbol{b}_{:k} \odot \boldsymbol{c}_{:k}\right]_+$ and $\boldsymbol{a}_{:k} = \boldsymbol{a}_{:k} / \|\boldsymbol{a}_{:k}\|^2$
    $\boldsymbol{b}_{:k} \leftarrow \left[\mathcal{X}_b^{(k)} \boldsymbol{a}_{:k} \odot \boldsymbol{c}_{:k}\right]_+$ and $\boldsymbol{b}_{:k} = \boldsymbol{b}_{:k} / \|\boldsymbol{b}_{:k}\|^2$
    $\boldsymbol{c}_{:k} \leftarrow \left[\mathcal{X}_c^{(k)} \boldsymbol{b}_{:k} \odot \boldsymbol{a}_{:k}\right]_+$
  end for
  $\mathcal{E} = \mathcal{X} - \sum_{k=1}^{K} \boldsymbol{a}_{:k} \circ \boldsymbol{b}_{:k} \circ \boldsymbol{c}_{:k}$
end while

III. NTF for Speech Enhancement

A. Feature

The aim of speech enhancement is to improve speech quality by removing or suppressing noises. Suppose speech $s(t)$ and noise $n(t)$ are two sound sources. The noisy speech $x(t)$ can be seen as the sum of speech and noise.

$$x(t) = s(t) + n(t)$$

t is a time index. Although this basic model is very simple, it is difficult to perform speech enhancement in the time domain alone. In most cases, it's safe to assume speech and noise are independent of each other. And short time Fourier transform (STFT) is adopted to generate a spectrogram. To take advantage of long term information, we group several adjacent time-frequency matrices in the spectrogram together to form a tensor, see Figure 2. Compared with the traditional way, the proposed feature extraction method is based on spectral features and has long term information.

B. Training

Let $\mathcal{S}$ and $\mathcal{N}$ denote speech and noise tensor features respectively. According to the above NFT factorization, the $\mathcal{S}$ and $\mathcal{N}$ can be factorized as

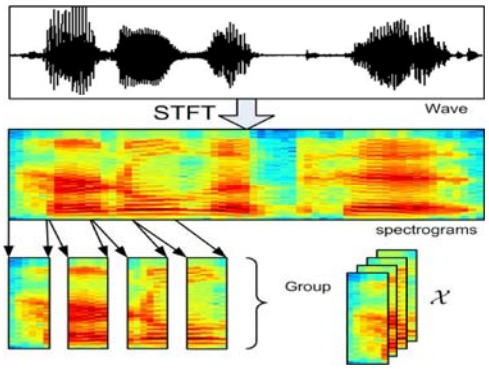

FIGURE II. NTF FEATURE EXTRACTION

$$\mathcal{S} = \sum_{k=1}^{K} \boldsymbol{a}_{s:k} \circ \boldsymbol{b}_{s:k} \circ \boldsymbol{c}_{s:k} + \mathcal{E}_s$$

$$\mathcal{N} = \sum_{k=1}^{K} \boldsymbol{a}_{n:k} \circ \boldsymbol{b}_{n:k} \circ \boldsymbol{c}_{n:k} + \mathcal{E}_n$$

where A_s and A_n, B_s and B_n are $R \times K$, $N \times K$ basis matrices which capture frequency and time structure of speech and noises respectively. C_s and C_n are $M \times K$ coefficient matrices. During the training phase, A_s and B_s are trained from the speech database and A_n and B_n are trained from the noise database by the algorithm 1.

C. Enhancement

The aim of enhancement is to separate the pure speech $s(t)$ from the noisy speech $x(t)$. After the training, the basis matrices of speech and noise are well estimated. We combine them together to estimate coefficient matrices of noisy speech.

$$A_x = [A_s, A_n]$$
$$B_x = [B_s, B_n]$$

Note that, both A_x and B_x have $2K$ vectors. The estimation equation is

$$\hat{\boldsymbol{c}}_{:k} = \mathcal{X}_{x,c}^{(k)} \boldsymbol{b}_{:k} \odot \boldsymbol{a}_{:k}, 1 \le k \le 2K$$

Here, $\mathcal{X}$ is the tenor formed by noisy speech. The first K vectors comprise the estimated coefficient matrix for speech and the rest comprises the estimated coefficient matrix for noise. The enhanced speech $\hat{\mathcal{S}}$ is obtained by the basis matrix and estimated speech coefficient matrix $\hat{\mathcal{S}} = \sum_{k=1}^{K} \boldsymbol{a}_{s:k} \circ \boldsymbol{b}_{s:k} \circ \hat{\boldsymbol{c}}_{s:k}$. And the time domain $\hat{s}(t)$ is separated from $x(t)$ by the inverse STFT of the $\hat{\mathcal{S}}$ subsequently.

D. Implementation

Overlapping is one of important issues in the implementation. In the process of tensor feature extraction, there are three steps involves segmentation. The first is the computation of STFT. The second is the grouping of adjacent spectral magnitude vectors to form a time-frequency matrix. And the last is the grouping of adjacent time-frequency matrices to form a tensor feature. Although the non-overlapping mode saves lots of computation and is easy for implementation, it often brings discontinuity at the boundary point and degrades the comfort of enhanced speeches. So we adopt 40% overlapping in each step from experimental results. During the phase of enhancement, the overlapping not only needs more computation but also requires some strategy to combine overlapping parts together. In our works, we take average strategy for its simplicity.

IV. EXPERIMENTS

A. Data

In our experiment, the speech data are from TIMIT database which consists of 630 speakers. Each speaker has 10 utterances. We select 5 utterances of each speaker as the training data and the rest as the test data. There are four types of noise data which are white Gaussian noise, bus noise, station noise and coach noise. The first type is simulated by a matlab code and the last three types of noise are recorded. Roughly speaking, the bus noise consists of the engine noise, occasional horning and automatic speech broadcasting. The station noise has the superposition of hundreds of person's voices and the automatic speech broadcasting. And the coach noise mainly contains the travelers' talks and the engine noise.

For each type of noise, we recorded at least 20 minutes, used 15 minutes for the training and preserved 5 minutes for the enhancement. These three types of noise can be seen as non-stationary noises. The noisy speech are produced by adding noise to speech at the SNR of -5, 0, 5 and 10 dB.

B. Configuration

The frame length and frame step is 400 and 160. The FFT order is 512. The $R = 256$, $N = 5$, $M = 5$ and $K = 50$ and iteration number is 200. Segmental signal to noise ratio (SSNR) and perceptual evaluation of speech quality (PESQ) are adopted as our evaluation tool. The classical SNR has a low correlation (~ 0.24) with the subjective assessment of speech quality and is not suitable for evaluating the estimated speech quality. The SSNR is defined as the average of SNR over segments with speech activity. It can quantify the level of non-stationary noise in speech in a better way and has a relatively high correlation (~0.77) with the subjective assessment of speech quality. Perceptual evaluation of speech quality (PESQ) calculates the distortion between the speech signal and the reference speech signal and gives a PESQ score based on the comparison result. The score is normalized between 0 and 4.5. The lower the score, the worse the quality.

C. Result

The result is shown in table.

TABLE I. SSNR AND PESQ OF ENHANCED SPEECH BY NTF

SSNR	white Gaussian noise		bus noise	
SNR(dB)	noisy	enhanced	noisy	enhanced
−5	−8. 773	−4. 128	−7. 719	−2. 5954
0	−7. 1359	−2. 1067	−5. 6475	−0. 4507
5	−4. 8644	−0. 0293	−2. 9801	1. 609
10	−2. 0383	2. 266	0. 152	3. 8796
SSNR	station noise		coach noise	
SNR(dB)	noisy	enhanced	noisy	enhanced
−5	−7. 4693	−2. 8045	−7. 7558	−2. 1221
0	−5. 3427	−0. 8975	−5. 7399	−0. 1691
5	−2. 6405	1. 1945	−3. 1222	1. 8182
10	0. 5353	3. 3232	−0. 0427	3. 8504
PESQ	white Gaussian noise		bus noise	
SNR(dB)	noisy	enhanced	noisy	enhanced
−5	0. 8724	1. 2069	1. 1411	1. 6507
0	1. 0746	1. 5921	1. 5162	1. 9842
5	1. 3546	1. 8754	1. 8926	2. 2414
10	1. 6967	2. 2155	2. 2294	2. 5482
PESQ	station noise		coach noise	
SNR(dB)	Noisy	enhanced	noisy	enhanced
−5	1. 0892	1. 5403	1. 3633	1. 8016
0	1. 3462	1. 8314	1. 7287	2. 099
5	1. 6578	2. 1437	2. 0759	2. 4151
10	1. 9909	2. 4213	2. 4064	2. 7171

D. Analysis

From the TABLE I, we conclude that both the SSNR and PSEQ are significantly improved by using the proposed method for all four types of noises. It means the NTF method is suitable not only for the stationary noise (white Gaussian noise) but also for the non-stationary noise (bus, station and coach noise).Similar to the ideal binary mask (IBM) in the computational auditory scene analysis (CASA) theory [14][14], the basis matrices in the NTF is more of a classifier rather than a filter. So they are capable of capturing long term and non-linear information which most STFT based enhancement algorithms fails in. We attribute this as the main reason for its good performance on the non-stationary signals. As we expected, the average improvement (average on different SNR) of SSNR and PESQ on the stationary noise (4.7035 dB and 0.4729) are slightly bigger than the average improvement on the non-stationary noise (4.5340 dB and 0.4130). The average improvement (average on different noisy types) of SSNR and PESQ at -5, 0, 5, and 10 dB are 5.0167, 5.0605, 4.5500, 3.6782 and 0.4334, 0.4603, 0.4237, 0.3947 respectively. With the increase of SNR, the speech enhancement task becomes harder, so the absolute improvement value becomes smaller. To our surprise, the best results is obtained at 0 dB rather than -5 dB. Perhaps, in the case of -5 dB, the noisy speech is too bad which may make the speech activity detection (VAD) and *etc*. algorithms inaccurate and lead to improper estimation.

V. CONCLUSIONS

In this paper, we propose an approach for speech enhancement based on the NTF. The proposed approach benefits from both the STFT and long term information. We adopt the HALS algorithm to factorize the feature tensors into basis matrices and coefficient matrices. Similar to the IBM in the CASA theory, the obtained basis matrices perform as a classifier more than a filter. This property makes it more suitable for non-stationary noisy speeches, which are the most common cases in application. The experiments were conducted on the TIMIT database and our collected real-life noise database. We evaluated the proposed method at different SNR and different types of noises. The experimental results demonstrated that the SSNR and PESQ are significantly improved (4.5764 dB and 0.4280 on average).

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant No. 61403224, 61370034 and 61273268.

REFERENCES

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.27, no.2, pp.113--120, 1979

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.32, no.6, pp.1109--1121, 1984.

[3] L. R. Rabine, Theory and Applications of Digital Speech Processing, Prentice Hall, 2010.

[4] G. Hinton, L. Deng, D. Yu and *et al.* “Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol.29, no.6, pp.82--97, 2012

[5] D. Yu, L. Deng and F. Seide, “The Deep Tensor Neural Network with Applications to Large Vocabulary Speech Recognition,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.21, no.2, pp.388--396, 2013.

[6] A. Cichocki, R. Zdunek, A. H. Phan and S. Amari, Nonnegative matrix and tensor factorizations, United Kingdom: John Wiley Sons, 2009.

[7] K. W. Wilson, B. Raj, and P. Smaragdis, “Regularized non-negative matrix factorization with temporal dependencies for speech denoising, ” Interspeech 2008 - 9th Annual Conference of the International Speech Communication Association, September 22--26, Brisbane, Australia, Processings, 2008, pp.411--414.

[8] N. Mohammadiha, T. Gerkmann, and A. Leijon, “A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization,” IEEE Workshop Applications of Signal Process, Audio Acoustics (WASPAA), October 6--9, New York, USA, 2011, pp.45--48.

[9] N. Mohammadiha, T. Gerkmann, and A. Leijon, “A new approach for speech enhancement based on a constrained nonnegative matrix factorization,” IEEE International Symposium on Intelligent Signal Process and Communication Systems (ISPACS), December, 7--9, Chiang Mai, Thailand, 2011, pp. 1--5.

[10] N. Mohammadiha, T. Gerkmann, and A. Leijon, “Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization, ” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.21, no.10, pp.2040 -- 2151, 2013.

[11] D. Fitzgerald, M. Cranitch, E. Coyle, “Non-negative Tensor Factorisation for Sound Source Separation, ” Proceedings of the Irish Signals and Systems Conference, Dublin, Ireland, 2005.

[12] T. Barker, T. Virtanen, “Non-negative Tensor Factorisation of Modulation Spectrograms for Monaural Sound Source Separation,” Interspeech 2013 -- 14th Annual Conference of the International Speech Communication Association, August 25--29, Lyon, France, Processings, 2013, pp.827--831.

[13] J. F. Gemmeke, T. Virtanen, A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol.19, No.7, pp.2067-2080, sep. 2011.

[14] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” In Divenyi P. (ed.), Speech Separation by Humans and Machines, pp. 181-197, Kluwer Academic, Norwell MA.

[15] K. Han, D. L. Wang. “A classification based approach to speech segregation,” Journal of the Acoustical Society of America, vol.132, pp. 3475--3483, 2012.