COMBINING NON-NEGATIVE MATRIX FACTORIZATION AND DEEP NEURAL NETWORKS FOR SPEECH ENHANCEMENT AND AUTOMATIC SPEECH RECOGNITION

Thanh T. Vu, Benjamin Bigot†, Eng Siong Chng

School of Computer Engineering, Nanyang Technological University, Singapore
†Rolls-Royce@NTU Corporate Lab, Nanyang Technological University, Singapore

ABSTRACT

Sparse Non-negative Matrix Factorization (SNMF) and Deep Neural Networks (DNN) have emerged individually as two efficient machine learning techniques for single-channel speech enhancement. Nevertheless, there are only few works investigating the combination of SNMF and DNN for speech enhancement and robust Automatic Speech Recognition (ASR). In this paper, we present a novel combination of speech enhancement components based-on SNMF and DNN into a full-stack system. We refine the cost function of the DNN to back-propagate the reconstruction error of the enhanced speech. Our proposal is compared with several state-of-the-art speech enhancement systems. Evaluations are conducted on the data of CHiME-3 challenge which consists of real noisy speech recordings captured under challenging noisy conditions. Our system yields significant improvements for both objective quality speech enhancement measurements with relative gain of 30%, and a 10% relative Word Error Rate reduction for ASR compared to the best baselines.

Index Terms— Speech Enhancement, Automatic Speech Recognition, Non Negative Matrix Factorization, Deep Neural Network, CHiME-3 challenge

1. INTRODUCTION

Speech enhancement (SE) aims to provide methods to improve the audio quality of noisy speech recordings. This topic has been studied for more than 50 years, and has produced successful approaches, especially statistics based methods [1] able to efficiently reduce the contribution of noise in degraded signals as long as the stationary noise assumption is respected. More recently, several works based-on machine learning algorithms such as Sparse Non-negative Matrix Factorization (SNMF) and Deep Neural Networks (DNN) have achieved significant improvements for non-stationary noises [2, 3].

SNMF-based SE methods [4], originated from [5], project the spectral features extracted from clean speech and noise signals into subspaces modelled as linear combinations of non negative basis vectors weighted by non negative activation coefficients. Enhancement of noisy speech is achieved in a supervised manner using the speech and noise basis vectors to estimate the speech and noise activation coefficients [3]. However, the linear mapping assumption used in SNMF will fail when speech and noise overlap in the feature domain or share similar bases. Several SNMF-based approaches have already addressed this limitation, first by jointly training the noise and speech basis vectors in order to produce more discriminant subspaces [6, 7, 8], and also by using non-linear mapping functions (typically with DNNs) to estimate the speech and noise coefficients [9].

DNN-based SE [2, 10] relies on the ability of Deep Neural Networks to estimate complex non-linear functions used to directly map log spectral features of noisy speech into corresponding clean speech signals and therefore may be more efficient in separating noise and speech in case of overlapping sub-domains. Temporal dependencies of speech are usually considered by extracting features on sliding context windows. DNN-based SE methods have been reported to yield good preservation of temporal and spectral speech characteristics. Nevertheless, training from raw speech features requires the estimation of billions of low-level parameters on potentially limited amount of data, and therefore may lead to poorly generic non-linear mapping functions.

In this paper, we propose a novel SNMF-based SE framework (presented Figure 1) integrating a Deep Neural Network. Contrarily to [9], the DNN is here used to produce a non linear mapping function between the SNMF activation coefficients of noisy signals to the equivalent activation coefficients of clean speech. One motivation of pre-processing noisy recordings with supervised NMF is that the projection of noisy signals into the lower dimension of NMF may first reduce the complexity of DNN training, and also produce a better DNN initialization, thanks to injecting prior knowledge gained with unsupervised SNMF on training data. In this work, we also propose a DNN architecture augmented by a supplementary layer in charge of reconstructing the log spectral features of the enhanced output speech signal. The injection of the reconstruction error of the output signal into the cost function of the machine-learning algorithm has previously been proven efficient using a discriminative SNMF-based framework during the estimation of the basis vectors from training data [6]. The reconstruction error computed as

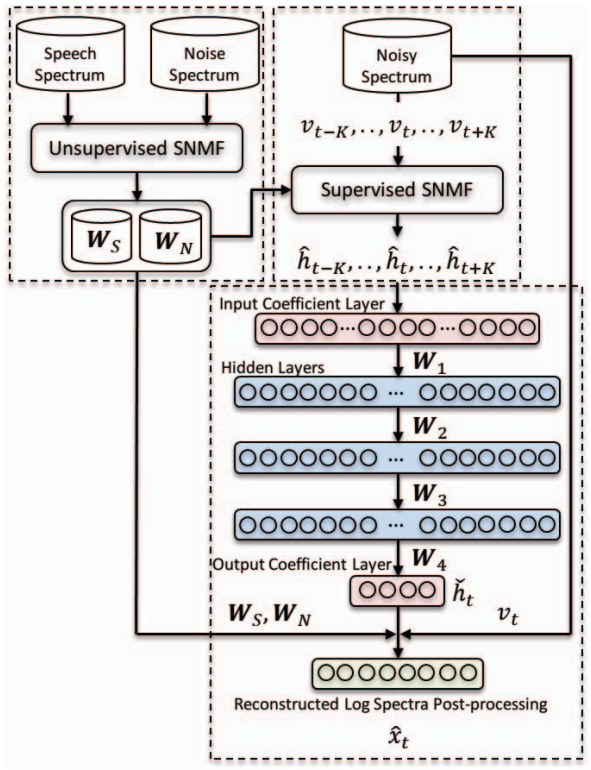

Fig. 1. Novel NMF & DNN-based Speech Enhancement

the distance between the log spectrum of the reconstructed signal and the log spectrum of the target clean speech is then back-propagated through the DNN at the learning time. Using this specific cost function, we expect to adapt the DNN to the final reconstructed signal considered in the evaluation of SE performances. Our proposal has been evaluated both for the tasks of Speech Enhancement and Automatic Speech Recognition (ASR) using several objective metrics. Our results have been systematically compared to several state-of-the-art DNN and SNMF-based SE systems [2, 3, 4, 9]. Evaluations have been conducted using the framework provided recently by the CHiME-3 challenge [11] on speech separation and recognition on challenging real noisy speech recordings. Evaluations yield that our proposal outperforms the state-of-the-art systems used for comparison on both SE and ASR tasks.

The remainder of this paper is as follows. In Section 2 we detail our novel SNMF-based SE framework employing DNN with modified cost function. Experiments are reported and discussed Section 3. We conclude this work in Section 4.

2. SYSTEM DESCRIPTION

We describe now a novel architecture derived from a SNMF-based Speech Enhancement framework (Figure 1). Our proposal consists in three main steps: an unsupervised learning of SNMF speech and noise basis vectors estimated on labelled data as in [4]; a supervised SNMF-based feature extraction from noisy speech recordings using the noise and speech bases estimated at the previous stage; a DNN-based SE module used to learn a non-linear mapping function between the SNMF activation coefficients and optimised to minimize the Mean Squared Error (MSE) between the log spectrum of the enhanced signal and the target clean speech.

2.1. NMF-based speech and noise bases estimation

We first estimate the basis vectors of clean speech and noise using the unsupervised SNMF algorithm [4]. SNMF assumes the spectral magnitude of a noisy signal $\mathbf{V} \in \mathcal{R}^{F \times T}$ (F the number of frequency bins and T the number of time frames) can be modelled as the linear combination of non negative basis vectors $\mathbf{W} \in \mathcal{R}^{F \times B}$ (with B the number of bases) and non negative activation coefficients $\mathbf{H} \in \mathcal{R}^{B \times T}$. The SNMF algorithm estimates $\mathbf{W}$ and $\mathbf{H}$ by minimizing the distance between $\mathbf{V}$ and $\mathbf{WH}$ computed using the Kullback-Leibler divergence and a sparseness constrain on $\mathbf{H}$ in the L_1 norm:

$$\mathbf{W}, \mathbf{H} = \min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}||\mathbf{WH}) + \mu||\mathbf{H}||_1 \quad (1)$$

$\mathbf{W}$ and $\mathbf{H}$ are estimated using iterative multiplicative update rules as described in [4].

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\overline{\mathbf{W}}^T \frac{\mathbf{V}}{\overline{\mathbf{WH}}}}{\overline{\mathbf{W}}^T \mathbf{1} + \mu} \quad (2)$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{\frac{\mathbf{V}}{\overline{\mathbf{WH}}}\mathbf{H}^T + \mathbf{1}(\mathbf{1}\mathbf{H}^T \odot \overline{\mathbf{W}}) \odot \overline{\mathbf{W}}}{\mathbf{1}\mathbf{H}^T + \mathbf{1}(\frac{\mathbf{V}}{\overline{\mathbf{WH}}} \odot \overline{\mathbf{W}}) \odot \overline{\mathbf{W}}} \quad (3)$$

We note $\mathbf{W}_S$ and $\mathbf{W}_N$ the bases of clean speech and background noise estimated on labelled training data. Experimentally, we applied 20 iterations of the algorithm, on single frame analysis windows and a sparseness constrain of 1.

2.2. Feature extraction using supervised SNMF

We first fix the speech and noise bases $[\mathbf{W}_S\, \mathbf{W}_N]$ estimated on training data, and then estimate the noise and speech activation coefficients $\hat{\mathbf{H}}_S$ and $\hat{\mathbf{H}}_N$ on noisy speech recordings using the iterative multiplicative update rules in equation 2.

The activation coefficients are then used as input features of the DNN, instead of raw spectral coefficients as in [9] or the log spectrum in [2]. For each frame of noisy speech (at index position t), we build a large vector composed of the concatenation of the activation coefficients of speech $\hat{h}_{S,t}$ and noise $\hat{h}_{N,t}$ vectors extracted on each frame on an analysis windows of width $(2K+1)$ frames centred on the t^{th} frame.

2.3. DNN training using SNMF-based reconstruction

A feed-forward DNN architecture presented Figure 1 is introduced to map noisy to clean activation coefficients. The

DNN consists of three sigmoid hidden layers and one sigmoid output layer. In order to obtain a more discriminative DNN training, we augmented the structure with one additional layer producing the reconstructed log spectrum vector $\hat{x}_t$ from the corresponding estimated NMF coefficients $\breve{h}_t = [\breve{h}_{S,t}^T \breve{h}_{N,t}^T]^T$ and v_t the input noisy spectral magnitude vector. We use the Wiener filter reconstruction as formalized in equation 4, with $\odot$ and $/$ the element-wise product and division.

$$\hat{v}_t(\breve{h}_t, \mathbf{W}_S, \mathbf{W}_N, v_t) = \frac{\mathbf{W}_S \breve{h}_{S,t}}{\mathbf{W}_S \breve{h}_{S,t} + \mathbf{W}_N \breve{h}_{N,t}} \odot v_t \quad (4)$$

$$\hat{x}_t(\breve{h}_t, \mathbf{W}_S, \mathbf{W}_N, v_t) = log\big(\hat{v}_t\big) \quad (5)$$

The objective function E to be minimized is the Mean Squared Error between the log spectrum of the reference x_t and reconstructed signals $\hat{x}_t$. The MSE is back-propagated to all layers of the DNN in a mini-batch training manner.

$$E = \frac{1}{2N} \sum_{t=1}^{N} \|x_t - \hat{x}_t\|_2^2 \quad (6)$$

The partial gradient of the cost function E used to estimate the network's weights can be expanded as follows:

$$\frac{\partial E}{\partial \mathbf{W}} = \frac{\partial E}{\partial \hat{x}_t} \frac{\partial \hat{x}_t}{\partial \breve{h}_t} \frac{\partial \breve{h}_t}{\partial \mathbf{W}} \quad (7)$$

We derive $\partial \hat{x}_t / \partial \breve{h}_t$ over speech coefficients $\breve{h}_{S,t}$ and noise coefficients $\breve{h}_{N,t}$ separately according to Equation 4. Using chain rule, these gradients can be derived as below:

$$\frac{\partial \hat{x}_t}{\partial \breve{h}_{S,t}} = \mathbf{W}_S \left[\frac{v_t \odot \big(s_t - r_t\big)}{\hat{v}_t \odot r_t^2} \right] \quad (8)$$

$$\frac{\partial \hat{x}_t}{\partial \breve{h}_{N,t}} = \mathbf{W}_N \left[\frac{v_t \odot s_t}{\hat{v}_t \odot r_t^2} \right] \quad (9)$$

where s_t and r_t are respectively:

$$s_t = \mathbf{W}_S \breve{h}_{S,t} \quad (10)$$

$$r_t = \mathbf{W}_S \breve{h}_{S,t} + \mathbf{W}_N \breve{h}_{N,t} \quad (11)$$

In the next section, we evaluate our proposal on both Speech Enhancement and Speech recognition tasks.

3. EXPERIMENTS

In the following section, our system will be denoted **(DNN-SNMF-Coef)**. Contrarily to [2, 9], where DNNs are trained of raw spectral features, we train the DNN on SNMF activation coefficients. Hence, to evaluate the influence of the input features of the DNN, we introduce a variant of our framework denoted **(DNN-SNMF-Spec)**, where the DNN is learned on spectral features to predict activation coefficients, and uses the modified cost function computed on signal reconstruction.

3.1. Data and Metrics

The dataset provided with the evaluation framework of the CHiME-3 challenge [11] on speech separation and recognition, is composed of real and simulated multi-channel noisy speech recordings captured in 4 challenging noisy environments: bus (BUS), cafeteria (CAF), pedestrian zone (PED) and street (STR). The training set is composed of 7138 utterances of read speech taken from the WSJ-0 corpus [12]. We prepare additional training data by simulating noisy speech with randomized Signal over Noise Ratio ($-5dB \leq SNR \leq +15dB$) using the tools provided by CHiME-3. Our evaluation set contains 2×2960 utterances (corresponding to the combined original DEV and TEST datasets of the campaign) for real and simulated noisy recordings.

Speech enhancement is evaluated in terms of Frequency-Weighted segmental SNR (fwSNRseg) [13] and Ceptrum distance (CEP) [14]. These metrics respectively measure the contribution of residual noise (fwSNRseg) and the speech distortion (CEP), and have both been reported to have high correlation with subjective test evaluations [13]. fwSNRseg measures the Signal over Noise Ratio between the weighted log power spectrum of clean target and the residual noise in the enhanced signal. The cepstrum distance CEP provides an estimate of the log spectral distance between two spectra. The performances on Automatic Speech Recognition are evaluated in terms of Word Error Rate (WER).

$$\text{WER}(\%) = \frac{\sum(\text{Insertion} + \text{Substitution} + \text{Deletion})}{\text{Nb of Words in reference}}$$

3.2. Baseline systems

Our system is systematically compared to several state-of-the-art NMF and DNN-based SE methods:

- **(SNMF)**: a conventional SNMF-based SE as in [3, 4];
- **(DNN)**: a DNN-based SE where the DNN maps directly noisy speech to clean speech as in [2];
- **(SNMF-DNN)**: a SNMF-based SE with DNN [9], mapping noisy speech spectrum to SNMF coefficients.

For every evaluated systems, the spectral features have been extracted with a Short Time Fourier with a $32ms$ Hamming weighting window and $8ms$-shift on signals sampled at $16kHz$. The dimension of the SNMF bases matrix is set to 257×100 (frequency bins x bases), estimated using 5% of clean WSJ-0 for $\mathbf{W}_S$ and 4×15 minutes of background noise (bus, cafeteria, street and pedestrian) for $\mathbf{W}_N$. The DNN is composed of 3 hidden layers of 3072 neurons. Its input features are extracted on a context window of 11 frames centred in the current frame. We follow the pre-training using Restricted Boltzman Machine [15] as described in [2], with a cross-validation $(90\% - 10\%)$ on simulated noisy speech for training and validation subsets. As a reminder, in the systems

(DNN), **(SNMF-DNN)** and **(DNN-SNMF-Spec)**, the DNN is trained on the spectral features. In our proposed system, **(DNN-SNMF-Coef)**, the DNN is trained with NMF activation coefficients vectors of dimension 200 for each frame (100 coefficients for speech and noise respectively).

The ASR system is a classical Hidden Markov Models with Gaussian Mixture Models acoustic models trained on the clean speech utterance of the WSJ-0 corpus, prepared using Kaldi [16] as described in the CHiME-3 ASR baseline [11].

3.3. Speech Enhancement and ASR Evaluation

Our system **(DNN-SNMF-Coef)** obtains the best performance for both fwSNRseg (7.59dB) and CEP (4.35) metrics as summarized Figure 2. It reached a gain of 7.59dB for fwSNRseg metric and outperformed the 3 baseline systems **(DNN)**, **(SNMF)** and **(SNMF-DNN)** with respective relative improvements of 160%, 80% and 31%. The **(DNN)** baseline performed surprisingly bad and produced less than 0.1dB improvement compared to the score measured on the raw noisy data. We assume this poor result of the **(DNN)** is caused by the nature of the fwSNRseg evaluation metric since the **(DNN)** performed well according to the CEP metric. We can see how the enhanced signal produced by the **(DNN)** contains a large contribution of residual noise but this DNN-based SE finally produced a relatively small distortion. We observe the benefit brought by introducing the modified cost function computed on the reconstructed enhanced signal by measuring the gain in performances obtained by **(DNN-SNMF-Spec)** against **(SNMF-DNN)**. The absolute improvement is equal to about 1.2dB in terms of fwSNRseg and 0.66 points of gain of CEP. By comparing the systems **(DNN-SNMF-Spec)** and **(DNN-SNMF-Coeff)**, we can appreciate how using either NMF activation coefficients or spectral features as input of the DNN impacts the performances. The absolute improvement is equal to 0.62dB in terms of fwSNRseg and 0.28 points of CEP. These promising results obtained by our system also highlight our approach is able to reduce both the contribution of residual noise and the level of distortion of the speech signal.

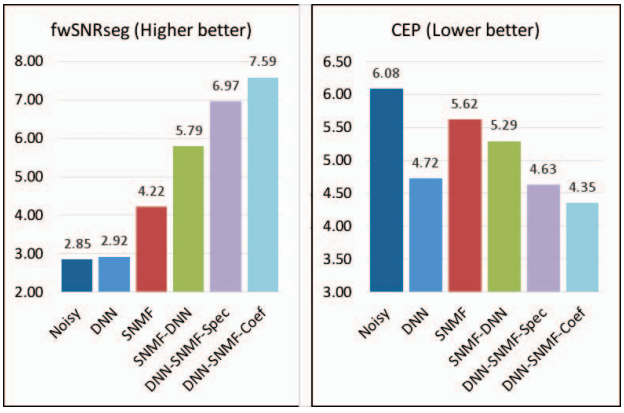

Fig. 2. Objective Evaluation with fwSNRseg and CEP metrics

Automatic Speech Recognition has been applied on the speech utterances enhanced by our methods and the baselines. For each enhancement method we report in Table 1 the overall WER obtained on real and simulated noisy speech utterances of the CHiME-3 test set. The **(DNN)** baseline speech enhancement improves significantly the WER with 47.6%. **(SNMF)** and **(SNMF-DNN)** systems improve WER by 6% and 19% respectively. Our proposed system achieves the best results with 43.7% WER, corresponding to 31% and 10% relative WER reduction compared to respectively the non-enhanced noisy speech and the **(DNN)** best baseline. Using NMF coefficients in **(DNN-SNMF-Coef)** or spectral features in **(DNN-SNMF-Spec)** as DNN inputs yields small difference in this experiment with 0.3%WER absolute improvement.

Table 1. WER (%) on simulated and real noisy speech

Speech Enhancement	Overall WER
No enhancement	63.0%
DNN	47.6%
SNMF	59.0%
SNMF-DNN	51.2%
DNN-SNMF-Spec	44.0%
DNN-SNMF-Coef	**43.7%**
Clean speech	**21.6%**

4. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a novel SNMF-based SE framework integrating Deep Neural Network. We trained DNNs to map SNMF activation coefficients of noisy speech to their clean version, by back-propagating the reconstruction errors of enhanced signals in the log spectral domain. Evaluations have been done on the real and simulated data of the CHiME-3 challenge and we have compared our proposal against several baseline methods. Our system has reached the best results by improving performances for both speech enhancement and Automatic Speech Recognition. Compared to the best baselines, we report a relative gain of 30% in terms of frequency-weighted segmental SNR, and 10% relative reduction of Word Error Rate. In future works, we will integrate more discriminative training of the SNMF bases and coefficients. We will also produce thorough analyses on the impact of SNMF and DNN parameters such as the architecture of DNN, the number of basis vectors and sparseness factor of the SNMF method.

5. ACKNOWLEDGEMENTS

This work was conducted within the Rolls-Royce@NTU Corp Lab with support from the National Research Foundation of Singapore under the Corp Lab@University Scheme.

6. REFERENCES

[1] P. C Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[2] Y.-H Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation based on improved Deep Neural Networks with dual outputs of speech features for both target and interfering speakers," in *Proc. ISCSLP*, 2014, pp. 250–254.

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *ICA'07*, 2007, pp. 414–421.

[4] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, pp. 88–101, 2008.

[5] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[6] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *ISCA Interspeech 2014*, 2014.

[7] Z. Wang and F. Sha, "Discriminative non-negative matrix factorization for single-channel speech separation," in *ICASSP 2014*, 2014, pp. 3749–3753.

[8] F. Weninger J. Le Roux, J.R. Hershey, "Deep NMF for speech separation," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2015, pp. 66–70.

[9] T.-G. Kang, K. Kwon, J.-W. Shin, and N.-S. Kim, "NMF-based target source separation using Deep Neural Network," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 229–233, 2015.

[10] Y. Xu, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, 2014.

[11] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[12] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete," in *Linguistic Data Consortium*, 2007.

[13] P. Loizou Y. Hu, "Evaluation of objective quality measures for speech enhancement," *IEEE Transaction on Speech Audio Processing*, vol. 16, no. 2, pp. 229–238, 2008.

[14] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 262–273, 1988.

[15] Y. Bengio, "Learning deep architectures for AI," *Foundat. and Trends Mach. Learn.*, vol. 2, pp. 1–127, 2009.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE.