INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A DNN-HMM Approach to Non-negative Matrix Factorization Based Speech Enhancement

Ziteng Wang, Xu Li, Xiaofei Wang, Qiang Fu, Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences
wangziteng@hccl.ioa.ac.cn

Abstract

General speaker-independent models have been used in non-negative matrix factorization (NMF) based speech enhancement algorithms for the practical applicability. And additional regulation is necessary when choosing the optimal models for speech reconstruction. In this paper, we propose a novel utilization of deep neural network (DNN) to select the models used for separating speech from noise. Specifically, multiple local dictionaries are learned, whereas only one is activated for each block in the separation step. Besides, the temporal dependencies between blocks are represented by hidden Markov model (HMM), with which it turns out a hybrid DNN-HMM framework. The most probable activation sequence is then solved by the Viterbi algorithm. Experimental evaluations which focus on a speech denoising application are carried out. The results confirm that our proposed approach achieves better performance when compared with some existing methods.

Index Terms: speech enhancement, non-negative matrix factorization, deep neural network, temporal continuity

1. Introduction

Non-negative matrix factorization (NMF) [1], [2] is a powerful approach for solving the speech enhancement problem. The basic idea is to decompose the source magnitude spectrum into a basis dictionary and a weight matrix, both of which are constrained nonnegative. The basis dictionary conveys meaningful dynamic patterns while the weight matrix represents the activation of different patterns along time [3].

Depending on the required training sources, NMF based methods are categorized into three classes: supervised, semi-supervised and unsupervised. Though the supervised methods prove quite effective, training samples of one or more sources are needed in advance. If no specific training data is provided, the unsupervised approach makes for a practical solution. The algorithm either learns dictionary online which relies on voice activity detection [4] or uses general speaker-independent examples [5]. For the latter case, several universal speech dictionaries are learned in order to describe the diverse spectral characteristics [6][7]. Since the local patterns should be mainly captured by some certain dictionary, block sparsity is introduced in the separation step. Each given frame is expressed as a linear combination of few optimum spectral vectors. Universal speech model (USM) [6] learns one separate dictionary for each speaker. Then only the ones that best fit the observed signal are activated. Sparsity is realized in a global sense by a penalty term of the weight coefficients. Mixture of local dictionaries (MLD) extends this concept [7]. It explores a more discriminative dictionary set which discovers the local convex cones. The block sparsity is temporally relaxed and the activation decision is made in the frame-by-frame way.

Another issue that comes up is the temporal correlations of speech, especially with the combination of multiple spectral components. Previous research reports that the utilization of this speech feature improves the final performance [8]. One popular practice to use the differences between the gains in adjacent frames as an additional cost term [9]. While some others suggest the usage of the Markov chain, for instance, factorial scaled hidden Markov model (FS-HMM) [10] and non-negative hidden Markov model (N-HMM) [11][12].

Meanwhile, the combination of NMF and an emerging technique of deep neural network (DNN) shows promising results [13][14]. In this paper, a hybrid framework of DNN and HMM is proposed, which furthers the idea of sparse modeling on the basis of multiple local dictionaries. Rather than imposing sparsity through a penalty term in the optimization criterion, we try to learn the optimal decision process from prior data. The motivation is related to two basic assumptions:

- The learned dictionaries well keep the local patterns so that only one dictionary is sufficient for recovering the spectral details of each block.
- If we know which dictionary to use, better reconstruction performance is expected. This process is realized with the utilization of DNN-HMM.

The paper is organised as follows. In Section 2 we give a review of the NMF based speech enhancement algorithm. Section3 is dedicated to our proposed method. The experiment setup and evaluation results are presented in Section 4. And in Section 5 we draw a concise conclusion.

2. NMF based speech enhancement

In the training phase, the short-time Fourier transform (STFT) magnitude spectrum $V \in R_+^{M\times N}$ of each isolated source is factorized into a dictionary matrix $W \in R_+^{M\times K}$ and a coefficient matrix $H \in R_+^{K\times N}$ by solving the optimization problem

$$\text{min. } D(V||WH) \quad \text{s.t. } W, H \geqslant 0 \tag{1}$$

where D is the Kullback-Leibler divergence or the square of Euclidean distance [2]. One general solution to the problem is

$$W \leftarrow W \odot \left\{ \left(\frac{V}{WH} \right) H^\top \right\} \tag{2}$$

$$H \leftarrow H \odot \left\{ W^\top \left(\frac{V}{WH} \right) \right\} \tag{3}$$

The product operator $\odot$ and division are performed in elementwise. $\top$ stands for transposition. W, H are randomly initialized and updated iteratively until convergence.

http://dx.doi.org/10.21437/Interspeech.2016-147

The learned dictionaries W_S for speech S and W_N for noise N are kept to reconstruct each source component in the enhancement phase. Assuming the additivity in the observed mixture, we have

$$V = S + N \tag{4}$$

$$= \begin{bmatrix} W_S & W_N \end{bmatrix} \begin{bmatrix} H_S \\ H_N \end{bmatrix} \tag{5}$$

If only one dictionary can be obtained, say W_S, the estimation of noise dictionary is implemented as

$$W_N \leftarrow W_N \odot \left\{ \left(\frac{V}{WH} \right) H_N^\top \right\} \tag{6}$$

When the source dictionary is of multiple spectral vectors, USM [6] and MLD [7] enforce sparse activations by the introduction of an extra penalty term $\Omega(H)$ into the optimization criterion (1). One more step is needed to shrink the coefficients

$$H \leftarrow \frac{1}{1 + \lambda/(\varepsilon + \|H\|_1)} H \tag{7}$$

where λ controls the amount of regulation and ε is a small number for numerical stability.

3. The DNN-HMM framework

An illustration of our proposed framework is shown in Figure 1. The observed mixture is represented by HMM with latent speech dictionaries as the hidden states. DNN is trained to give prediction of state posterior probabilities. The first-order Markov assumption further constrains the transition possibility between states. Before training, clustering is performed on clean speech examples to generate the desired dictionary set and provide training labels for DNN. Here the DNN-HMM part is much different from the phoneme dependent NMF as proposed in [15] that relies on a separate robust speech recognizer, which itself is a challenging task.

Dictionaries for reconstructing the target source are identified with the most probable state sequence $\boldsymbol{S} := \{s_t\}$, which is obtained by HMM decoding given parameters $\{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$. $\boldsymbol{\pi}$ is the initial state probability vector. $\boldsymbol{A} := \{a_{ij}(s_t = j | s_{t-1} = i) \mid \forall i, j\}$ denotes the transition matrix and $\boldsymbol{B}$ indicates the emission density matrix. Detailed processing procedures are presented in the following subsections.

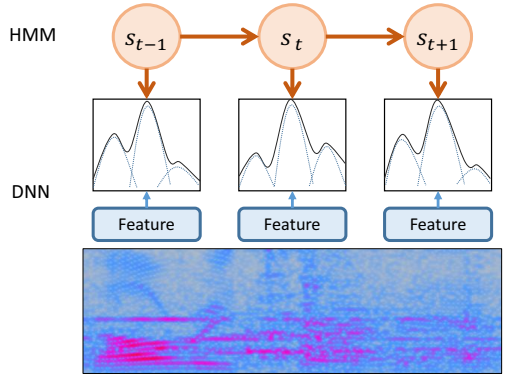

Figure 1: *Illustration of the proposed framework.*

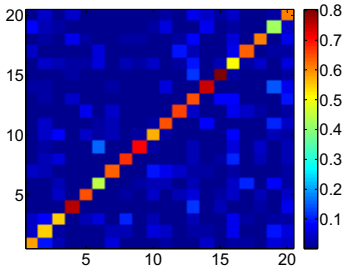

Figure 2: *An example of state transition matrix between blocks for a total cluster number of 20.*

3.1. Clustering

Given clean training source data, the purpose of clustering is to discover the diverse spectral characteristics. Spectral vectors in the same cluster are close in the space manifold and tend to possess similar structure. Here k-means clustering is used for its simplicity while other methods are also feasible. Spectrogram similarity is measured by the Euclidian Distance. One dictionary for each cluster is learned using equation (2)(3). Suppose the cluster number is C, we have C local dictionaries in total.

Considering the dictionary as a latent state of each cluster and labeling each spectrum block with its cluster number, a first-order HMM is applied to describe the signal. The state transition probability is naturally calculated from all the training samples

$$a_{ij}(s_t = j | s_{t-1} = i) = \frac{\sum_t (s_t = j | s_{t-1} = i)}{\sum_t (s_{t-1} = i)} \tag{8}$$

where $s_t = i$ means that block t belongs to the ith cluster, and $i, j \in [1, \ldots, C]$. An example of the transition matrix is depicted in Figure 2. Temporal correlations show up as a tendency that consecutive blocks stay in the same state, which is proved by the larger diagonal elements in the matrix.

The prior probability of each state is

$$\boldsymbol{\pi}(s_0 = i) = \frac{\sum_t (s_t = i)}{T} \tag{9}$$

where T is the number of total training blocks.

3.2. Deep neural network

The clean speech is mixed with different kinds of noise to provide prior training data for DNN. The network is supposed to classify each mixture block into the *right* speech cluster. Training labels are binary vectors that use the 1-of-C coding scheme with all elements being zeros except for element c. Multiclass cross-entropy criterion is chosen as the cost function

$$Er = \sum p \log(p) + (1 - p) \log(1 - p) \tag{10}$$

The network architecture consists of L hidden layers and all layers are fully connected. ReLU activation components are used for the hidden layers and Softmax component is used for the output layer.

Each layer is first pre-trained as a restricted Boltzman machine (RBM). Back-propagation technique of conjugate gradient descent (CGD) is then applied for fine tuning the network. The error on a validation set is checked for choosing the model of the best performance.

During testing, the outputs of DNN are interpreted as the probabilities of one block belonging to each states.

3.3. Viterbi decoding

Provided with the above parameters, the Viterbi algorithm is applied to find a most likely state sequence that maximizes the function

$$P(\boldsymbol{V}, \boldsymbol{S}) \approxeq \max_{\boldsymbol{S}} \boldsymbol{\pi}(s_0) \prod_t a(s_t|s_{t-1}) p(\boldsymbol{v_t}|s_t) \tag{11}$$

where $\boldsymbol{V} := \{\boldsymbol{v_t}\}$ is the observation. The emission probability is obtained using the Bayes's theorem

$$p(\boldsymbol{v_t}|s_t) = \frac{p(s_t|\boldsymbol{v_t}) p(\boldsymbol{v_t})}{p(s_t)} \tag{12}$$

in which the posterior density comes from DNN and the denominator is exactly the state prior probability. The final states are retrieved by back-tracking.

3.4. Reconstruction

The dictionaries $W = [W_1,\ W_2, ..., W_t]$ for speech reconstruction is solely decided by the state sequence. Meanwhile, there are two ways to utilize this information. We can either solve the separation problem in an online mode using the one speech dictionary for each block, or settle it in a batch mode by concatenating the C individual dictionaries together and activating just the *right* coefficients. We use the second mode in latter experiments for its better performance. After random initialization of H_S, the other coefficients are set to be zeros and remain zeros throughout except for the ones corresponding to the *right* dictionary. So no extra regularization is needed in the process.

The ultimate estimation of speech spectrogram is computed by a Wiener-type filtering [5]

$$\tilde{S} = \frac{W_S H_S}{W_S H_S + W_N H_N} \odot V \tag{13}$$

The time-domain signal is recovered by inverse STFT using the phase of the mixture and overlap-and-add method.

4. Evaluation

4.1. Experiment setup

Our proposed approach is tested in a speech denosing application. The experimental implementations largely follow that of USM [6] and MLD [7] for comparison. All signals are sampled at 16 kHz. STFT is performed using 1024 samples with 75% overlap. 20 male speakers (10 sentences each) are randomly chosen from the training set of TIMIT corpus to provide general data for learning speech dictionaries. Each dictionary holds $R_S = 10$ basis vectors. The number of clusters is $C = 20$. Clean files of 5 speakers from the test set are mixed with 10 non-stationary noise signals, namely {bird, casino, cicadas, computer keyboard, eating chips, frogs, jungle, machine guns, motorcycles, ocean} [16], at 0dB signal-to-noise ratio (SNR), which makes a total of 50 sentences for testing.

The parameters of DNN are $L = 2$ hidden layers each with 512 nodes. The logarithm magnitudes of one single frame is used as input, so the feature size turns out 513. Noise segments different from testing are mixed with clean speech, which results in 2000 training examples. 20% files are selected as the validation set.

In the unsupervised scenario, we apply the universal speech dictionary while the noise dictionary is learned from the mixture. The size of the noise dictionary is chosen depending on its type [16], i.e. one of $\{20, 10, 200, 20, 20, 10, 20, 10, 10, 10\}$.

The BSS toolbox [17] and the perceptual evaluation of speech quality (PESQ) [18] are used for quality measurement. PESQ is highly correlated to human speech quality scores while signal-to-distortion ratio (SDR) serves as an overall separation performance metric.

4.2. Results and analysis

We compare the proposed method with the classical Log-MMSE algorithm [19], USM and MLD. The results are shown in Table 1. The USM and MLD algorithms are tuned to best performance based on the setup in respective references. DNN-D refers to the case that we take the state of maximum probability as the final state and leave out the Viterbi search.

Table 1: *SDR(dB) and PESQ (bottom) scores of different approaches averaged on all test sentences.*

Noisy	[19]	USM	MLD	DNN-D	Proposed
0.00	3.37	6.42	8.28	8.94	9.07
1.81	1.90	1.92	2.17	2.36	2.38

The usage of DNN indicates a distinct improvement over baseline methods. The increment in SDR over MLD is about 0.7 dB and 0.19 in PESQ. The consideration of temporal continuity results in a further enhancement and scores as high as 9.07 dB. Note that the dictionaries in MLD are different from that of DNN-D, though both involve some clustering techniques. The difference does not contribute to its final performance when we replace the dictionary set of MLD.

The achievement of our proposed approach largely lies in the introduced prior information by deep neural network. With the same types of noise for training, DNN is able to recognize the *right* speech cluster from the mixture. Spectral details are then recovered by the corresponding dictionary. Rather than imposing sparsity through a penalty term, the proposed method basically decides the optimal dictionaries to be activated. The point is verified from another perspective by an *oracle* test, in which the mixtures for training the network are put to test. Since we already know the true speech labels, *oracle* results are obtained. The SDR averaged on all the 200×10 files is 9.37 dB. This serves as an important reference. Somehow, DNN suffers certain performance loss in mismatched test conditions.

To illustrate the effects of temporal constraints, the spectrograms of one test sentence are presented in Figure 3. The speech is corrupted with computer keyboard noise. As is shown by this figure, the utilization of HMM produces more smooth spectrum. In the last subgraph, less fluctuations in dictionary states are observed between adjacent frames of the proposed method than of DNN-D. In this way, there is a reduced distortion in speech active regions.

Next, we further discuss our proposed method in two aspects, of which the results are in Table 2.

Table 2: *Performance comparison in the semi-supervised and strict unsupervised case. SDR(dB)*

	USM	MLD	DNN-D	Proposed
semi-	10.37	10.31	10.50	10.56
un-	6.42	8.28	7.58	7.71

Semi-supervised (semi-): Since we have got noise samples for training the neural network, the noise dictionary could be

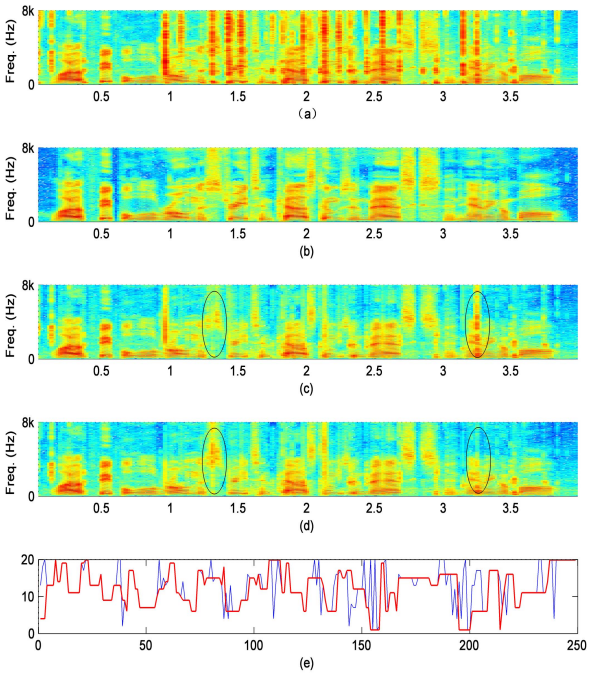

Figure 3: Spectrograms of noisy signal at 0dB SNR (a), of clean signal (b) , of DNN-D output (c) and of the proposed method (d). The bottom graph (e) records the predicted state/dictionary labels for each frame. The blue line corresponds to DNN-D and the red line refers to the DNN-HMM based method.

directly learned from the sample segments, while the third-part speech models remain unchanged. This is compared to semi-supervised separation of the baseline algorithms. All methods are adjusted with regard to iteration and the size of noise dictionary for optimal performance. The DNN-D reaches 10.50 dB and the DNN-HMM based approach scores 10.56 dB, both of which are slightly better than the baseline.

Strict unsupervised (un-): No noise samples of the test data are available beforehand. Instead, general noises are exploited to train the network, which confirms with a strict unsupervised setup. We choose the 100 nonspeech environmental sounds set [20] for its wide coverage and the difficulty of our evaluation task. Unexpectedly, there is a dramatic drop of about 1.4 dB in the scores. Though it is still 1.3 dB better than USM, for which the discriminative dictionaries might be the main reason. We conclude that in the unmatched noise conditions misclassification disturbs the final outcome. Nevertheless, supplementary experiments see better results with increased noise types for training.

5. Conclusions

An endeavour has been made to integrate the prior knowledge learned from prior data into NMF based speech enhancement algorithms. The supervised technique of DNN is employed for the optimal model decision in the case of multiple alternative dictionaries. With the combination of HMM, the consequent hybrid framework takes both the spectral dynamics and the temporal continuity of speech into consideration. Its effectiveness is confirmed in a certain speech denoising experiment.

We note that the generalization of the neural network part remains a issue. And future work also includes the exploration of representative dictionary set, long temporal dependencies and relaxation of using one dictionary per block.

6. Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 61271426, 11504406, 11590770, 11590771, 11590772, 11590773, 11590774), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500, XDA06040603), National 863 Program (No. 2015AA016306), National 973 Program (No. 2013CB329302) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

7. References

[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[3] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using non-negative factorizations: A unified view," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 66–75, 2014.

[4] M. N. Schmidt, J. Larsen, and F. T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *IEEE Workshop on Machine Learning for Signal Processing*, Aug 2007, pp. 431–436.

[5] A. L. Nasser Mohammadiha, Paris Smaragdis, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[6] D. L. Sun and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 141–145.

[7] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *Signal Processing Letters, IEEE*, vol. 22, no. 3, pp. 293–297, 2015.

[8] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising." in *Interspeech*, 2008, pp. 411–414.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 121–124.

[11] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Latent variable analysis and signal separation*. Springer, 2010, pp. 140–148.

[12] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 17–20.

[13] D. S. Williamson, Y. Wang, and D. Wang, "Deep neural networks for estimating speech model activations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[14] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "Nmf-based target source separation using deep neural network," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 229–233, 2015.

[15] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent nmf for speech enhancement in monaural mixtures." in *INTERSPEECH*, 2011, pp. 1217–1220.

[16] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online plca for real-time semi-supervised source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 34–41.

[17] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] I. Rec, "P. 862.2 wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union, CH–Geneva*, 2005.

[19] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[20] G. Hu, "100 nonspeech environment sounds [online]." Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2004.