RECOGNIZE AND SEPARATE APPROACH FOR SPEECH DENOISING USING NONNEGATIVE MATRIX FACTORIZATION

Fahad Sohrab, Hakan Erdogan

Faculty of Engineering and Natural Sciences Sabanci University Istanbul, Turkey
{fahadsohrab,haerdogan}@sabanciuniv.edu

ABSTRACT

This paper proposes a novel approach for denoising single-channel noisy speech signals. A speech dictionary and multiple noise dictionaries are trained using nonnegative matrix factorization (NMF). After observing the mixed signal, first the type of noise in the mixed signal is identified. The magnitude spectrogram of the noisy signal is decomposed using NMF with the concatenated trained dictionaries of noise and speech. Our results indicate that recognizing the noise type from the mixed signal and using the corresponding specific noise dictionary provides better results than using a general noise dictionary in the NMF approach. We also compare our algorithm with other state-of-the-art denoising methods and show that it has better performance than the competitors in most cases.

Index Terms— Speech enhancement, single channel denoising, nonnegative matrix factorization.

1. INTRODUCTION

Everywhere in our surroundings different types of noises exist which may greatly degrade the quality of sound in communication systems. These background signals can be of various types and natures. Noises at crowded places such as airports, restaurants and markets have different spectro-temporal characteristics. It is crucial to denoise the mixed signal, particularly when the receiver's end uses automated systems for voice recognition.

In [1] an enhancement algorithm called tracking of non-stationary noise based on MMSE estimation of speech spectral amplitudes is proposed. Standard approaches such as spectral subtraction [2] and Wiener filtering [3] are restricted to denoising of stationary noise only as they require signal and/or noise estimates for processing. Method for performing source separation using general training data is proposed in [4].

For an improved denoising system, the method should adapt itself to the noise type. In [5] general formalism for source model adaptation in the framework of Bayesian models is introduced while in [6] phoneme-dependent NMF based algorithm for separating speech from monaural mixtures is presented. A denoising system that makes use of the knowledge of the noisy environment type can perform much better than a generic denoising system.

In this work, we propose a *"recognize and separate"* approach to denoise a noisy signal. In this approach, we first recognize the type of noise present in the noisy signal and then we use nonnegative matrix factorization [7] as the basic algorithm for speech enhancement. The main goal is to separate the desired speech signal from the background noise signal. NMF is already used for separating a speech signal from different types of source signals like another speech signal or music [8,9]. By using NMF we can model different types of background noise signals.

The rest of the paper is organized as follows. In Section 2 we explain the basic NMF algorithm, in Section 3 problem formulations is presented. In Section 4, the process of training the dictionaries using NMF and training a classifier is explained. The trained dictionaries are later used in the testing phase for denoising. In Section 5 we explain the basic testing procedure in our method. We introduce our experimental results in Section 6. Finally, we present our conclusions in Section 7.

2. NONNEGATIVE MATRIX FACTORIZATION

Nonnegative matrix factorization factors the nonnegative n-by-m matrix V into a nonnegative basis (or dictionary) matrix B (n-by-k) and a weight matrix W (k-by-m).

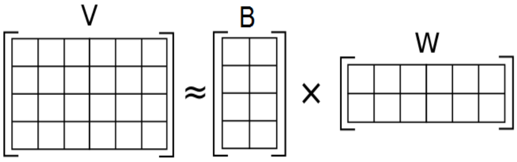

Fig. 1. Illustration of NMF.

$$V \approx BW$$

The decomposition BW usually cannot be exactly equal to V, so some cost functions are used for penalizing the difference in NMF. In [10] Kullback-Leibler (KL) divergence between V and BW was found to work well for audio source separation. We will restrict ourselves to KL divergence given below.

$$D(V \| BW) = \sum_{i,j} \left(V_{i,j} \log \frac{V_{i,j}}{(BW)_{i,j}} - V_{i,j} + (BW)_{i,j} \right)$$

We will work in magnitude spectral domain so the non-negativity constraint is necessary.

$$B, W \geq 0$$

A solution for minimizing this cost function can be found by iteratively updating B and W using the following equations.

$$B \leftarrow B \otimes \frac{\frac{V}{BW} W^T}{1W^T} \quad (1)$$

$$W \leftarrow W \otimes \frac{B^T \frac{V}{BW}}{B^T 1} \quad (2)$$

The matrix $B = [b_1, b_2, b_3 ... b_k]$ is the nonnegative "dictionary matrix" whose column explains the columns of $V = [v_1, v_2, v_3 ... v_m]$ matrix as follow

$$v_i \approx \sum_{j=1}^{k} w_{ij} b_j .$$

Where w_{ij} are nonnegative as well.

3. PROBLEM FORMULATION

In a single channel denoising problem the observed mixed signal is the addition of speech and noise signals.

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

We apply NMF in the magnitude spectra domain so after observing the signal we take its short time Fourier transform and then consider its absolute value. Thus we can approximate our mixed signal's magnitude spectrogram X as

$$X = S + N .$$

Now the goal is to first recognize the noise type present in the mixed signal and then use the corresponding dictionary matrix in NMF formulation to extract the speech signal S from the mixed signal X as we explain in the following sections.

4. TRAINING STAGE

4.1 Training the dictionaries

We used NMF for training the dictionary of noises. To train the noise and speech dictionaries, we first take magnitude of the STFT of the given training data and then we use (1) and (2) to find the basis for speech and specific noise signals.

$$S_{train} \approx B_{speech} W_{speech}$$

$$N_{train} \approx B_{noise} W_{noise} ,$$

where B_{speech} and W_{speech} denote the dictionary and weight matrices of speech while B_{noise} and W_{noise} denote the dictionary and weight matrices of noise. We keep B_{speech} and B_{noise} as speech and noise dictionaries and discard W_{speech} and W_{noise} which are not used in our approach. For general noise dictionary, we used a single audio file obtained by concatenating all the noise files used for experiments.

4.2 Training the noise classifiers

Training a classifier is the process of feeding the known data belonging to a specified known class and creating a classifier on the basis of that known data. There are various classifiers available which can be useful for classification of unknown data to determine its class.

We train frame-level noise classifiers using known noise data for each noise type we consider in this work. The features used for classification are magnitude spectral vectors obtained from the corresponding noise data. We normalize the spectral vectors by dividing by their l_2 norm in order not to get affected by energy differences in training and test stages. One can also use more complex features such as MFCCs or log-filterbank features which are used in speech recognition. However we found that these simple normalized magnitude spectrum features worked satisfactorily for our purposes in this work.

For recognition of noise type we trained and tested different classifiers such as linear Bayes normal classifier, k-nearest neighbor classifier, naive Bayes classifier and quadratic Bayes normal classifier. All the classifiers were trained using pure noise data. The details are discussed in Section 6

5. TESTING STAGE

For testing we first mix a noise signal with a speech signal which we recorded. All the noises were mixed with clean speech file at various target SNRs. The parts of noise signals used for training were not used for mixing. Also random portions of noise were selected for mixing with clean speech. After identifying the type of noise present in mixed noisy signal we again use NMF but this time we kept the concatenated basis fixed for decomposing the mixed signal magnitude spectrogram X as follows

$$X = \left[B_{speech,} B_{noise} \right] W .$$

The speech is estimated by multiplying the bases matrix with its corresponding weight matrix

$$S_e = B_{speech} W_s ,$$

$$N_e = B_{noise} W_n ,$$

where S_e is the estimated speech spectrogram and W_s is the sub matrix inside W which corresponds to the speech dictionary.

$$W = \begin{bmatrix} W_s \\ W_n \end{bmatrix}.$$

After estimating the speech spectrogram, we used the Wiener filter for soft masking [8].

$$\widetilde{S} = H_{wiener} \otimes S,$$

where

$$H_{wiener} = \frac{S_e^{\,2}}{S_e^{\,2} + N_e^{\,2}},$$

and $\otimes$ is the a binary operation that takes two matrices of the same dimensions, and produces another matrix where each element is the product of elements of the original two matrices while S_e and N_e represent estimated speech and noise respectively.

$\widetilde{S}$ is the magnitude STFT of the reconstructed signal. We used the phase of the mixed signal's STFT and use inverse STFT to obtain the denoised signal in the time domain.

6. EXPERIMENTS, RESULTS AND DISCUSSION

In our experiments we used four types of noises for evaluating our algorithm and prove its versatility. We considered fan noise, factory noise, water tap noise and machine gun noise for our experiments. We recorded fan noise and water tap noise ourselves, while factory noise and machine gun noise were taken from the Aurora2 database [12]. We also recorded a 10 minute long single person speech signal for training the speech dictionary. All the data of our experiments can be found at [13]. We used a total of 24 mixed signals for our experiments.

For our experiments we used a 12 second long audio file for training a classifier for each type of noise. We initially test the classifier accuracies on a held out noise data of 2.5 seconds. Table 1 show the train and test error on different classifiers at frame level.

Classifier	Train Error	Test Error
Linear Bayes Normal	0.75	0.75
K-Nearest Neighbor	0.00	0.08
Naive Bayes	0.19	0.22
Quadratic Bayes Normal	0.26	0.27

Table 1. Train and test errors of classifiers.

The least error we achieved was by using the k-nearest neighbour (KNN) classifier so we opted for KNN as the classifier. We used prtools toolkit for matlab [11] for classification in our experiments.

For testing the separation performances, we used a single speech utterance mixed with four different types of noise at SNR values from -10 to 15 dB in 5 dB increments. In our experimental setup we achieved good results by using 16 dictionary entries for speech and 4 dictionary entries for noise. For evaluating and comparing our results with state-of-the-art techniques we used most common performance measures i.e. SDR [14], SIR [14], SNR and PESQ [15]. We used audio files sampled at 8000 Hz.

STFT of the signals is obtained by performing fast Fourier transform of Hamming windowed frames of 100 samples where each successive frame is shifted by 50 samples. In the mixed signal we searched for the lowest energy frames that comprise 10% of the whole signal and classified them using the k-nearest neighbor classifier and combine the decision for each frame by score averaging to recognize the noise type in the utterance. We got 84 percent accuracy in recognizing the noise type correctly from the mixed signal at various SNRs.

In Tables 2, 3, 4 and 5 we are reporting SDR, SIR, SNR and PESQ performance respectively for each method at various input SNRs. Each column in every table represents a different type of method used for denoising. "Noise Tracker" is the method proposed in [1] and "Wiener-AS" is the Wiener filter with apriori SNR estimation method proposed in [16]. "NMF general" refers to the method where a general noise dictionary is used for denoising and "NMF specific" refers to the proposed "recognize and separate" method where we use a noise-specific dictionary for denosing. The results reported for "NMF specific" include the overall performance numbers for 16 percent misrecognized files as well. The performance numbers are averages over 4 different noise types.

DB	Noise Tracker	Wiener-AS	NMF general	NMF specific
-10	-7.03	-7.77	-8.99	-5.28
-5	-1.42	-2.21	-2.28	0.63
0	3.69	2.83	1.97	4.44
5	8.47	7.45	6.52	8.99
10	12.99	12.01	11.01	12.28
15	17.57	16.57	14.23	14.05

Table 2. Source to Distortion Ratio (SDR) after denoising the mixed signal using different algorithms.

DB	Noise Tracker	Wiener-AS	NMF general	NMF specific
-10	29.88	32.17	26.57	40.66
-5	37.28	38.51	34.07	38.32
0	45.50	45.13	42.78	45.81
5	48.64	51.33	46.55	50.62
10	55.22	53.83	51.24	55.64
15	59.51	57.38	50.03	54.13

Table 3. Source to Interferences Ratio (SIR) after denoising the mixed signal using different algorithms.

DB	Noise Tracker	Wiener-AS	NMF general	NMF specific
-10	-4.58	-4.82	-6.82	-3.07
-5	-0.05	-0.44	-1.02	2.07
0	4.30	3.81	2.90	5.46
5	8.73	8.04	7.08	9.55
10	13.10	12.20	11.27	12.55
15	17.62	16.01	14.34	14.12

Table 4. Signal to Noise Ratio (SNR) after denoising the mixed signal using different algorithms.

DB	Noise Tracker	Wiener-AS	NMF general	NMF specific
-10	1.70	1.60	1.68	1.85
-5	2.02	1.99	2.03	2.09
0	2.42	2.36	2.33	2.42
5	2.66	2.60	2.53	2.60
10	2.85	2.81	2.72	2.82
15	3.05	3.02	2.93	3.09

Table 5. Perceptual Evaluation of Speech Quality (PESQ) after denoising the mixed signal using different algorithms.

From Tables 2, 3, 4 and 5 in some cases "Noise Tracker" is found to work better for mixed signals of high SNR values i.e. 10 and 15 DB but at low SNR values the proposed "NMF specific" approach is performing better than all other competitors. On average "NMF specific" is found to work well in every case.

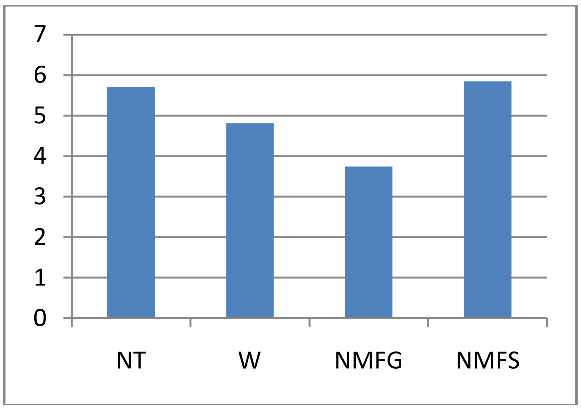

Fig.2. Average of Source to Distortion Ratio.

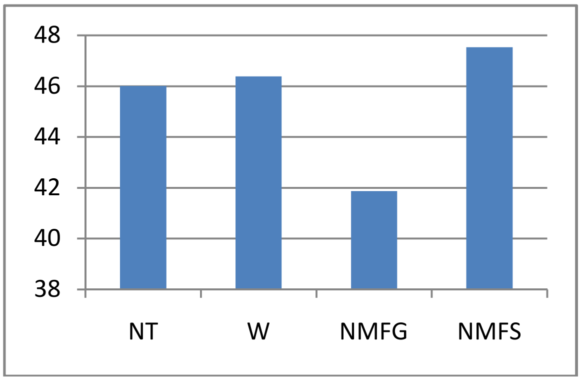

Fig.3. Average of Source to Interferences Ratio

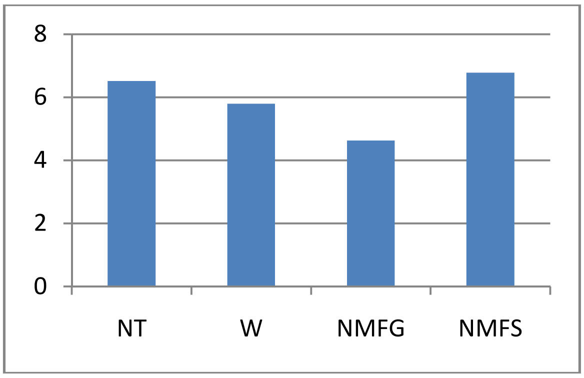

Fig.4. Average of Signal to Noise Ratio.

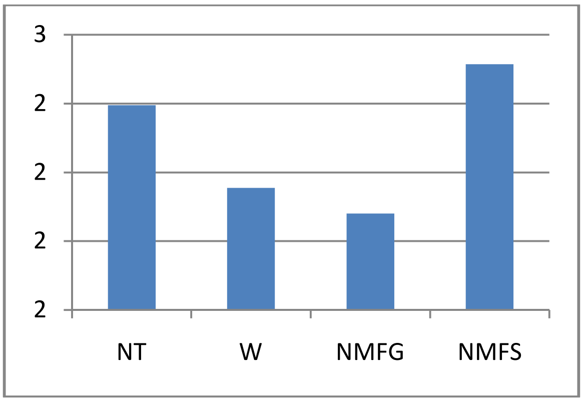

Fig.5. Average of Perceptual Evaluation of Speech Quality.

Figure 2, 3, 4, 5 illustrate the average values across the columns of Tables 2, 3, 4 and 5 respectively. On average, "NMF specific" (NMFS) appears to be the best approach among the considered alternatives when all possible input SNR values are considered.

7. CONCLUSION

In this work we studied the enhancement of noisy speech signals by identifying the noise type in the signal. We experimentally showed that in specific environments of noise, NMF performs better than algorithms designed for generic noise. In the future it can be further improved by considering a speaker specific approach by training dictionaries for speakers. There may be more room for research in this area where the dictionaries can be trained from the mixed signals directly.

8. REFERENCES

[1] Erkelens, Jan S., and Richard Heusdens. "Tracking of nonstationary noise based on data-driven recursive noise power estimation." *Audio, Speech, and Language Processing, IEEE Transactions on* 16.6 (2008): 1112-1123.

[2] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." *Acoustics, Speech and Signal Processing, IEEE Transactions on* 27.2 (1979): 113-120.

[3] Lim, Jae S., and Alan V. Oppenheim. "Enhancement and bandwidth compression of noisy speech." *Proceedings of the IEEE* 67.12 (1979): 1586-1604.

[4] Sun, Dennis L., and Gautham J. Mysore. "Universal speech models for speaker independent single channel source separation." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[5] Ozerov, Alexey, et al. "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs." *Audio, Speech, and Language Processing, IEEE Transactions on* 15.5 (2007): 1564-1578.

[6] Raj, Bhiksha, Rita Singh, and Tuomas Virtanen. "Phoneme-Dependent NMF for Speech Enhancement in Monaural Mixtures." *INTERSPEECH*. 2011.

[7] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," Advances in *Neural Information Processing Systems*, vol. 13, pp. 556–562,2001.

[8] Grais, Emad M., and Hakan Erdogan. "Single channel speech music separation using nonnegative matrix factorization and spectral masks." *Digital Signal Processing (DSP), 2011 17th International Conference on*. IEEE, 2011.

[9] Grais, Emad M., and Hakan Erdogan. "Regularized nonnegative matrix factorization using Gaussian mixture priors for supervised single channel source separation." *Computer Speech & Language* 27.3 (2013): 746-762.

[10] Wilson, Kevin W., Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran."Speech denoising using nonnegative matrix factorization with priors." *ICASSP*. 2008.

[11] Duin, R., et al. "PRTools 4.1." *A Matlab Toolbox for Pattern Recognition, Software and Documentation downloaded May* (2010).

[12] Hirsch, Hans-Günter, and David Pearce. "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions." *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*. 2000.

[13] Fahad Sohrab. "Speech (NMF) Data." Sabanci university *http: //myweb.sabanciuniv.edu/fahadsohrab/research/*

[14] Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation." *Audio, Speech, and Language Processing, IEEE Transactions on* 14.4 (2006): 1462-1469.

[15] Hu, Yi, and Philipos C. Loizou. "Evaluation of objective quality measures for speech enhancement." *Audio, Speech, and Language Processing, IEEE Transactions on* 16.1 (2008): 229-238.

[16] Scalart, Pascal. "Speech enhancement based on a priori signal to noise estimation." *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings. 1996 IEEE International Conference on*. Vol. 2. IEEE, 1996.