Speech Enhancement Using Convolutive Nonnegative Matrix Factorization with Cosparsity Regularization

Majid Mirbagheri[1], *Yanbo Xu*[1], *Sahar Akram*[1], *Shihab Shamma*[1,2]

[1]Institute for System Research, University of Maryland College Park, USA
[2]Department of Electrical and Computer Engineering, University of Maryland College Park, USA
mbagheri@umd.edu, yanbohsu@umd.edu, sakram@umd.edu, sas@umd.edu

Abstract

A novel method for speech enhancement based on Convolutive Non-negative Matrix Factorization (CNMF) is presented in this paper. The sparsity of activation matrix for speech components has already been utilized in NMF-based enhancement methods. However such methods do not usually take into account prior knowledge about occurrence relations between different speech components. By introducing the notion of cosparsity, we demonstrate how such relations can be characterized from available speech data and enforced when recovering speech from noisy mixtures. Through objective evaluations we show our proposed regularization improves sparse reconstruction of speech, especially in low SNR conditions.

Index Terms: speech enhancement, cosparsity, convolutive nonnegative matrix factorization

1. Introduction

Enhancing quality and intelligibility of noisy signals by reducing amount of noise in them has been an active field of research in the recent decades. Numerous approaches have been developed to address this problem [1]. Among them, methods based on Nonnegative Matrix Factorization (NMF) has gained interest due to their ability to handle nonstationary noises present in real-world applications. Standard NMF first introduced by Lee and Seung [2] simply aimed to approximate a non-negative matrix $X \in \mathbb{R}_{\geq 0}^{M \times L}$ as the product of two nonegative matrices $W \in \mathbb{R}_{\geq 0}^{M \times R}$ and $H \in \mathbb{R}_{\geq 0}^{R \times L}$ where $R \leq M$. With X being a magnitude spectrogram, NMF performs a linear basis decomposition storing the basis functions (atoms) in the columns of W and their corresponding temporal evolutions (activations) in the rows of H. Standard NMF ignores potential dependencies across successive columns of X. In order to account for temporal context CNMF was introduced by Smaragdis [3] by extending approximation of X as following:

$$X \approx \sum_{\tau=0}^{T-1} W(\tau) \overrightarrow{H}^{\tau} \tag{1}$$

where $\{W(\tau)\}$ is a set of time-varying bases, $W(\tau) \in \mathbb{R}_{\geq 0}^{M \times R}$ and $H \in \mathbb{R}_{\geq 0}^{R \times L}$ contains the activities. The operator $\overrightarrow{(.)}^{\tau}$ performs time-shifting by zero-padding of its operand with τ columns of zeros to the left and truncating that at the right to maintain correct dimensionality. Usually the approximation is done by solving a constrained optimization problem in which a divergence function between the input and its approximation is tried to be minimized subject to the nonnegativity of the constructing matrices. To measure the reconstruction error, in this work we used the Frobenious norm, $\|.\|_F$, (i.e., the square root of summed squared matrix entries).

$$\arg\min_{W(\tau),H} D(X\|\hat{X}) \text{ subject to } W(\tau), H \geq 0 \quad \forall\tau. \tag{2}$$

Now assuming additivity in the magnitude spectra domain, NMF-based speech enhancement methods usually aim to have each basis function (atom) in the final decomposition of the mixture only describe the speech or the noise spectrograms. In this way, enhancement would simply be achieved by combining speech components according to their corresponding activities in the mixture.

We should point out the fact that the additivity assumption about the magnitude spectra of the speech and noise, i.e. $X = S + N$, does not generally hold, however it has been shown that it is acceptable for the goal of source separation [3, 4].

Chickoki in [5] introduced a general NMF framework in which speech and noise separation in decomposition was achieved by means of regularizing the structure of the bases and their activations with regularization terms that penalize the reconstruction error in (2). Such regularizations usually take into account statistical characteristics, e.g. independence, or prior knowledge about the representation of the signal in hand (i.e. speech). A well-known example for the latter case is the sparsity of activations in CNMF representation of clean speech. It has been observed that preserving sparsity of speech activations usually results in better separation of speech and non-speech components [4] especially in presence of wideband noise.

A common measure to quantize sparsity is $\ell 1$-norm of speech activations over time. One issue about this and in general other sparsity measures used in this context is that they are only useful to minimize the global sparseness of the representation without accounting for how the occurrence of different components of clean speech are mutually correlated to each other. The importance of incorporating such information becomes specifically highlighted when the noise components resemble those of speech (e.g. second talker or babble noise) and hence are susceptible to activating speech bases. Wilson in [6] took advantage of prior information by assuming a normal distribution with known parameters for both speech and noise activations. In addition to being noise-dependent, one major shortcoming of this method was that their assumption implicitly mandated a certain value for both speech and noise signal powers. To incorporate prior information about activations without facing such issues, in this work, we introduce an extended notion of sparsity ,namely cosparsity. Having this measure quantize relative activation of bases pairs, the new regularization on activations forces the components that are cosparse in clean speech not to

co-occur in the denoised segment. We will show through some experiments how the new regularization can improve estimation of the speech spectrogram, $\hat{S}$, when used along with the standard sparsity term. In the following two sections, we first explain the cosparsity and the corresponding penalty function and then describe the regularized-CNMF method based on that.

2. Cosparsity

We define *cosparsity* between the activations of i-th and j-th components at the time instance l, in the following manner:

$$c_{ij}^{l} = \frac{h_{il}^2 + h_{jl}^2}{h_{il} h_{jl}} \tag{3}$$

with h_{il} being the activation of the i-th component at the time instance l. Note that for nonnegative activations the cosparsity measure always takes nonnegative values greater than or equal to 2. It takes its minimum value of 2 when activations are equal and approaches infinity as the ratio between the activations gets bigger and bigger or simply when the two components are cosparse. Note that the cosparsity is a symmetric relation and being only a function of relative strength, its value does not vary by scaling activations.

We shall maintain different levels of cosparsity among all pairs of components according to their record of cosparsity learned from an available clean speech dataset. This is done by prioritizing having larger cosparsity between pairs that are seldomly active at the same time in clean speech. We assume that there is a speech corpus available for training that can be used to learn the prior knowledge on speech components. A basis, $\widetilde{W}_S(\tau)$, and the corresponding activation matrix, $\widetilde{H}_S(\tau)$ are pretrained on the speech corpora using standard CNMF.

We use the codebook activations, $\widetilde{H}_S$, to calculate a matrix P, keeping track of cosparsity of component pairs. In order to maintain the high cosparsity for pairs with corresponding low entries in P, we minimize a regularization term. This new term is basically the sum of the bounded inverse of cosparsity measure, $\frac{h_{il}h_{jl}}{h_{il}^2+h_{jl}^2}$, for all pairs weighted by the entries in P. Entries in P are between 0 and 1, where a value close to 1 reflects a cosparse pair and one close to 0 occurs when the activations are very similar. Having these properties in mind, we formed the entries in P as:

$$p_{ij} = (1 - \frac{\tilde{\mathbf{h}}_i . \tilde{\mathbf{h}}_j}{|\tilde{\mathbf{h}}_i||\tilde{\mathbf{h}}_j|})^{\zeta} \tag{4}$$

with $\tilde{\mathbf{h}}_i$ being the i-th row in the matrix $\widetilde{H}_S$ and $\tilde{\mathbf{h}}_i.\tilde{\mathbf{h}}_j$ and $|\tilde{\mathbf{h}}_i|$ respectively being the inner product and the $\ell 2$-norm of the vectors. The constant ζ simply controls the distribution of the entries of P on the interval $[0, 1]$. High values of ζ enforces cosparsity on smaller number of pairs with very high records of cosparsity while a very low value enforces cosparsity to all the pairs evenly.

3. CNMF with cosparsity regularization

Given a noisy speech spectrogram $X \in \mathbb{R}_{\geq 0}^{M \times n}$, the proposed regularized CNMF forms the estimate spectrogram as following:

$$\widehat{X} = \sum_{\tau=0}^{T-1} \left[W_S(\tau)\ W_N(\tau) \right] \begin{bmatrix} \overrightarrow{H_S}^{\tau} \\ \overrightarrow{H_N}^{\tau} \end{bmatrix} \tag{5}$$

Here, $W_S(\tau) \in \mathbb{R}_{\geq 0}^{M \times R_S}$ and $W_N(\tau) \in \mathbb{R}_{\geq 0}^{M \times R_N}$ respectively denote speech and noise bases while $H_S \in \mathbb{R}_{\geq 0}^{R_S \times L}$ and $H_N \in \mathbb{R}_{\geq 0}^{R_N \times L}$ represent their activations. The estimate spectrogram $\widehat{X}$ is then computed by minimizing the following cost function with respect to the nonnegative basis and activation matrices:

$$\mathcal{J} := \frac{1}{2}\|X - \widehat{X}\|_F + \alpha \cdot J_S(H_S) + \beta \cdot J_C(H_S) \tag{6}$$

where $J_S(H_S)$ and $J_C(H_S)$ respectively represent sparsity and cosparsity regularization terms computed as:

$$J_C(H_S) = \sum_{i \neq j} p_{ij} \sum_{l=1}^{L} {c_{ij}^{l}}^{-1} \tag{7}$$

and

$$J_S(H_S) = \sum_{l=1}^{L} \|\mathbf{h}_S^l\|_1 \tag{8}$$

with $\|\mathbf{h}_S^l\|_1$ denoting the $\ell 1$-norm of the l-th column in H_S and P being the cosparsity penalty matrix. α and β are two constants determining the amount of punish for the sparsity and cosparsity terms. A higher value for the constant α yields in lower value/number of active components for the speech part, and for the constant β results in a generally higher degree of cosparsity maintained between components. Similar to standard CNMF, the optimization will be performed through initializing the entries in each of these matrices and then a series of alternating updates on the basis and activation matrices according to multiplicative rules [7]. In order to preserve the nonnegativity of these matrices, this procedure updates them by gains that are a function of the terms in the corresponding gradient of the cost function $\mathcal{J}$. For any of these matrices say A, considering that the partial derivative matrix of the objective function with respect to elements in A can be decomposed into two nonnegative parts as:

$$\left.\frac{\partial \mathcal{J}}{\partial A'}\right|_{A'=A_{\text{old}}} = \nabla^+ - \nabla^- \tag{9}$$

In this way, we will have the multiplicative update rule as:

$$A_{\text{new}} = A_{\text{old}} \odot \frac{\nabla^-}{\nabla^+} \tag{10}$$

where $\odot$ is the Hadamard product (element-wise multiplication), and division between the matrices is also an element-wise operation. For the cosparsity term, these two nonegative parts ∇_C^+ and ∇_C^- can be element-wise derived as:

$$\frac{\partial J_C}{\partial h_{il}} = \sum_{j;i\neq j} \frac{p_{ij} h_{jl}^3}{(h_{jl}^2 + h_{il}^2)^2} - \sum_{j;i\neq j} \frac{p_{ij} h_{jl} h_{il}^2}{(h_{jl}^2 + h_{il}^2)^2} = \delta_{il}^+ - \delta_{il}^- \tag{11}$$

Thus following [5], the new multiplicative update for speech activations would be expressed by:

$$H_S \leftarrow \left\langle H_S \odot \frac{W_S^T(\tau) \overleftarrow{X}^{\tau} + \nabla_C^-}{W_S^T(\tau) \overleftarrow{\widehat{X}}^{\tau} + \beta \cdot \mathbf{1}_{R_S \times L} + \nabla_C^+} \right\rangle_{\tau} \tag{12}$$

Since noise activations only appear in the error term of $\mathcal{J}$ the multiplicative update for them would be:

$$H_N \leftarrow \left\langle H_S \odot \frac{W_S^T(\tau) \overleftarrow{X}^{\tau}}{W_S^T(\tau) \overleftarrow{\widehat{X}}^{\tau}} \right\rangle_{\tau} \tag{13}$$

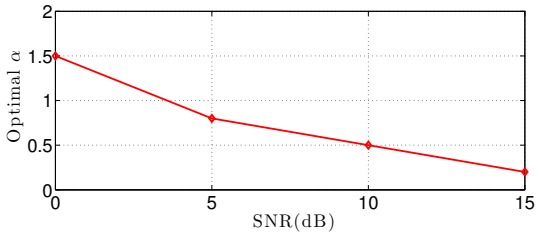

Figure 1: *Optimal value for the sparsity term weight α computed for pink noise at different SNR levels.*

Following the same procedure, the basis matrix $W(\tau) = \left[W_S(\tau)\ W_N(\tau)\right]$ is also updated in the following way:

$$W(\tau) \leftarrow W(\tau) \odot \frac{X \overset{\tau\rightarrow T}{H}}{\widetilde{X} \overset{\tau\rightarrow T}{H}} \tag{14}$$

with $H = \begin{bmatrix} H_S \\ H_N \end{bmatrix}$. We also normalize the basis columns after each multiplicative update so that they all have their $\ell 1$-norms equal to one. The entries in the basis are initialized by the pre-trained codebook, $\widetilde{W}_S$. It should be noted that the initialization is necessary not only for its known importance in optimization but also as kind of a labeling of speech components whose pairs are supposed to have uneven degrees of cosparsity. All other three matrices are initialized with random nonnegative values. The alternation between updates on the basis and activations is continued until either the relative change in $\mathcal{J}$ is lower than 1% or the number of iterations exceeds 150 (whichever happens first). The speech spectra is then simply estimated as:

$$\widehat{S} = \sum_{\tau=0}^{T-1} W_S(\tau)\ \overset{\tau\rightarrow}{H}_S \tag{15}$$

Using the phase info from the noisy spectra and the overlap-add method, $\angle X$, we then generate the enhanced waveforms.

4. Results and Experiments

In order to assess the performance of our proposed method, we used a speech corpora consisting utterances from the TIMIT database. The speech waveforms all sampled at their original rate, 16 kHz, were transformed to magnitude spectrogram (short-time Fourier transform) using 32 ms Hamming-weighted windows overlapped by 50%. The speech codebook and the cosparsity penalty coefficients were computed on about three minute of clean speech from randomly-chosen male and female speakers in TIMIT `train` subset. For the testing, we used 20 sentences from 10 male and 10 female speakers from TIMIT `test` subset. These sentences were corrupted by additive noise at four different SNR levels ranging from 0 dB to 15 dB. Three different noise types chosen from NOISEX dataset were evaluated in our experiments, *Pink* (a stationary noise with energy uniformly spread over log frequency scale), and *City* and *Babble* as two cases of nonstationary noise.

Speech qualities were measured by a standard objective metric, Perceptual Evaluation of Speech Quality (PESQ) [8]. The measure was particularly developed to model subjective tests commonly used in telecommunications. However it has been commonly used in assessing quality of speech enhancement algorithms as well. PESQ takes values between .5 (bad) and 4.5 (no distortion).

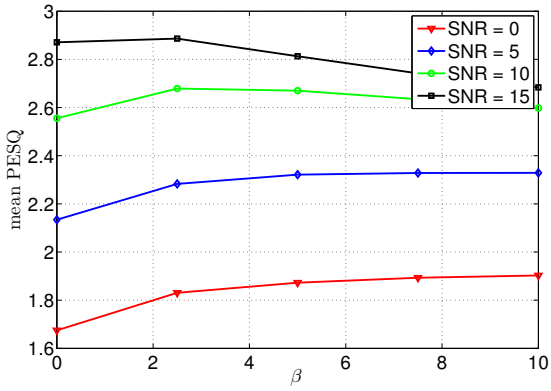

Figure 2: *mean PESQ vs. cosparsity term weight, β.*

In the experiments, we set $R_S = 100$, $R_N = 50$, and $T = 3$. In order to investigate how the quality of enhancement is effected by selection of the parameters α, β and ζ, the algorithm was run on corrupted samples in pink noise at different SNR levels using different combinations of these parameters. We considered $\alpha \in [0, 5]$, $\beta \in [0, 10]$ and $\zeta \in [0.1, 100]$.

For the sparsity term constant, α, the optimal value was computed by averaging the PESQ scores across speakers and different values of the two other parameters. For each SNR level, the value giving rise to the maximum quality was chosen as the optimal one. Figure 1 demonstrates these optimal values of α calculated for four different SNR levels. This result is consistent with the one reported in [4], and confirms that in lower SNRs, a more sparse reconstruction of speech results in a better quality. Having set α to its optimal values for each SNR condition and averaging PESQ scores over different values of β, we observed that a value roughly equal to 20 for the parameter ζ almost always resulted in the highest scores. Finally, using these optimal values found for α and ζ, we looked at the average PESQ for different values of β. Figure 2 shows how the selection of β affects quality of reconstructed speech segments. It is observed that for all SNR levels except 15dB the maximum improvement is achieved for a value of β greater than zero suggesting the effectiveness of having cosparsity regularization term in the CNMF framework. Similar to the trend for the parameter α, it is observed here that as the intensity of noise is increased, the optimal value of beta also increase. This means that enforcing cosparsity relations between speech components becomes more essential as the noise gets stronger and able to activate speech components more frequently.

We finally compared our proposed method against the regular sparse CNMF and a baseline speech enhancement method. For the baseline method we chose a spectral subtraction algorithm introduced in [9]. The comparison we did was best versus best, i.e. for the regular sparse CNMF, we simply set β to zero, and reported the highest mean PESQ score over all values of the parameter α while for ours it was the highest score over all values of α and all non-zero values of the parameter β ($\zeta = 20$). The results of the comparison are illustrated in figure 3. The improvement with respect to regular sparse CNMF is obvious for all three noise types especially in lower SNRs. Our method outperforms the baseline one for *Pink* and *City* noise. However, the results for *Babble* noise are somehow weaker. We believe this is related to speech-like statistical properties of this type

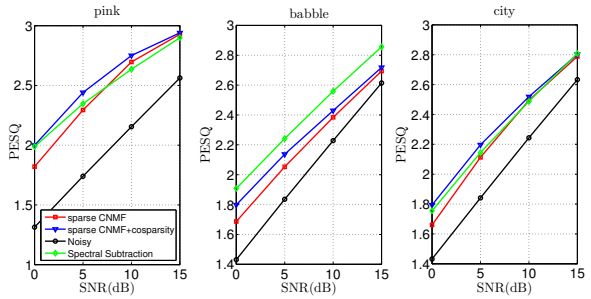

Figure 3: *PESQ Improvements for different noise types.*

of noise which poses a challenge to methods based on *a priori* knowledge of speech.

5. Conclusions

In this paper, we proposed a new regularization for enhancement of noisy speech signals based on a relation between activations of components pairs, denoted by *cosparsity*. We discussed how selection of parameters can impact the quality of the reconstructed speech signals, and showed through objective evaluations that our proposed algorithm can effectively improve quality of enhancement especially in low SNR conditions.
It should also be noted that the cosparsity regularization technique introduced in this paper can naturally be incorporated to any application dealing with part-based decompositions of input signals.

6. References

[1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC, 2007, vol. 30.

[2] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[3] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, 2007.

[4] R. De Fréin and S. T. Rickard, "Learning speech features in the presence of noise: Sparse convolutive robust non-negative matrix factorization," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–6.

[5] A. Cichocki, R. Zdunek, and S.-i. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5. IEEE, 2006, pp. V–V.

[6] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4029–4032.

[7] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2858–2864, 2009.

[8] T. ITU and P. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Recommendation ITU-T*, 2001.

[9] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech communication*, vol. 50, no. 6, pp. 453–466, 2008.