A Speech Dereverberation Method Using Adaptive Sparse Dictionary Learning

*M. Moshirynia, F. Razzazi, A. Haghbin**

Department of Electrical and Computer Engineering
IAU, Science & Research Branch
Tehran, Iran
m.moshirinia@sri.ac.ir, {razzazi, haghbina*}@srbiau.ac.ir

P.1.2

ABSTRACT

We present a monaural blind dereverberation method based on sparse coding of deconvolved version of reverberated speech signal in a dictionary which is learned by joint dictionary learning method, consisting of the concatenation of a clean speech and a non-negative matrix factor deconvolution result of the reverberated copy. The environment specific dictionary is originally learned off-line on a training corpus for different locations, while adaptive dictionary learning continues on-line for any other surroundings. Our approach uses both non-negative blind deconvolution and sparse coding, and achieves some improvements on objective voice quality testing's like perceptual evaluation of speech quality.

PROPOSED METHOD

Our structure is based on the sparse and non-negative nature of magnitude of speech signals in STFT domain as recently proposed for speech enhancement. We assume that the phase of the reverberated signal can approximate phase of the dereverberated signal as is commonly used in the derivation of speech enhancement algorithms e.g. spectral subtraction, adaptive filtering, and subspace approach.

Motivations

Convolutive reverberation can be expressed in terms of RIR. Equivalently, in STFT domain we may write:

$$S_r[n] = \sum_m S_c[m] H[n-m] \tag{1}$$

where S_r, S_c and H are magnitudes of reverberated signal, clean signal and RIR in frequency domain respectively. Equation (1) can also be written in the form of operators as following:

$$S_r \cong \hat{S}_r = \sum_{m=1}^{M} \hat{S}_{c,m} H^{m\rightarrow} \tag{2}$$

where the arrow operator shifts the columns of its argument by *m* spots to the right.

Non-negative Matrix Factor Deconvolution (NMFD): beginning with the decomposition of non-negative matrix into multiplication of two non-negative matrices and where $P < M$ such that we minimize the error of reconstruction of V by $W \cdot H$ using the cost function introduced by Lee *et al.* [17]:

$$D = \left\| V \otimes \ln\left(\frac{V}{W.H}\right) - V + W.H \right\|_F \tag{3}$$

which yields to an iterative solution:

$$\begin{cases} H = H \otimes \dfrac{W^T \dfrac{V}{W.H}}{W^T.1} \\ W = W \otimes \dfrac{\dfrac{V}{W.H} H^T}{1.H^T} \end{cases} \tag{4}$$

where the cross operator enclosed by a circle is Hadamard product (an element-wise multiplication) and divisions are element-wise too. Setting

$$\Lambda = \sum_{t=0}^{T-1} W_t H^{t\rightarrow} \tag{5}$$

and using cost function

$$D = \left\| V \otimes \ln\left(\frac{V}{W.H}\right) - V + W.H \right\|_F \tag{6}$$

results:

$$\begin{cases} H = H \otimes \dfrac{W_t^T \left(\dfrac{V}{\Lambda}\right)^{\leftarrow t}}{W_t^T.1} \\ W_t = W_t \otimes \dfrac{\dfrac{V}{\Lambda} H^{t\rightarrow T}}{1.H^{t\rightarrow T}} \end{cases} \tag{7}$$

substituting W_t, V and Λ with $\hat{S}_{c,m}$, S_r and $\hat{S}_r$ consequently yields:

$$\begin{cases} H = H \otimes \dfrac{\hat{S}_{c,m}^T \left(\dfrac{S_r}{\hat{S}_r}\right)^{\leftarrow m}}{W_t^T.1} \\ \hat{S}_{c,m} = \hat{S}_{c,m} \otimes \dfrac{\dfrac{S_r}{\hat{S}_r} \left(H^{m\rightarrow}\right)^T}{1.H^{m\rightarrow T}} \end{cases} \tag{8}$$

Joint Dictionary Learning: In a pioneering work on producing supper resolution images by Yang *et al.* [16], a pair of jointly learned dictionaries (D_b, D_s) is used, one dictionary for blurred samples and the other for sharp samples. During training, a dictionary D is learned to represent both sharp and blurred examples simultaneously with the same sparse code e.g. a; then D is split into two distinct dictionaries and to represent blurry and sharp samples in consequence. At test time, given a new blurry sample x, a sparse code a is obtained by decomposing x using D_b, and ones hopes to be a good estimate of the unknown sharp sample.

There is an interesting relationship between dictionary learning method used for image processing application to extract supper resolution samples from the blurry one and our application which aims to enhance the speech spectrogram of deconvolved version of reverberated speech signal, $\hat{S}_c$, which is sparse enough and will have possibly overcomplete dictionary.

Having both clean and reverberated speech signals at training time, we first deconvolve reverberated signal to get $\underline{\hat{S}_c}$ and then use it as a training set similar to what is done in deblurring application as blurry patches. Therefore, we may write:

$$S_j \cong D_j \alpha \tag{9}$$

or

$$\begin{bmatrix} S_c \\ \hat{S}_j \end{bmatrix} \cong \begin{bmatrix} D_c \\ D_r \end{bmatrix} \alpha \tag{10}$$

where D_j, D_c and D_r are named joint, clean and reverberated dictionaries respectively.

Proposed Architecture

Our approach is based on two distinct steps: deconvolution and enhancement. Deconvolution step is commonly applied in both training and test, but enhancement has different story. For the enhancement step, a possibly overcomplete dictionary of atoms is trained jointly using joint dictionary learning, for clean and deconvolved version of reverberated copy of speech magnitudes which are then split into two distinct dictionaries named as clean and reverberated. In the enhancement step, an observation of reverberated speech is first deconvolved and then sparsely coded in the reverberated dictionary. The clean speech magnitude is estimated by multiplying clean dictionary to the extracted sparse code. This estimate is combined with the post-processed phase of the reverberated signal to produce the time domain signal.

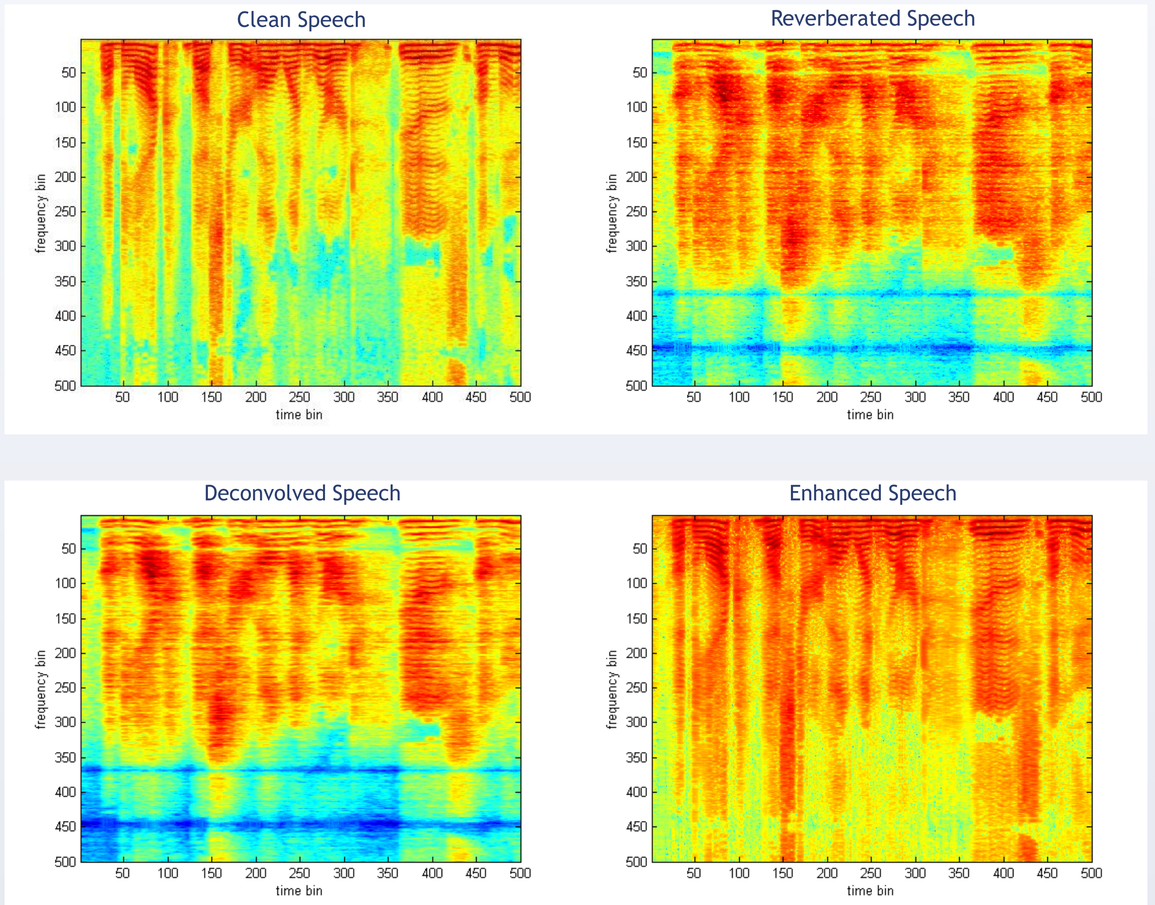
We supposed that the observed reverberated speech magnitude is the convolution result of clean speech magnitude and RIR. The goal of the deconvolution step is to obtain an estimate of clean speech and an estimate of the RIR. For the formal analysis, we distinguish between convolutive and non-convolutive reverberation effects (e.g. classroom and studio, respectively), and make use of results from sparse coding theory to enhance only reverberated speech in the convolutive environments.

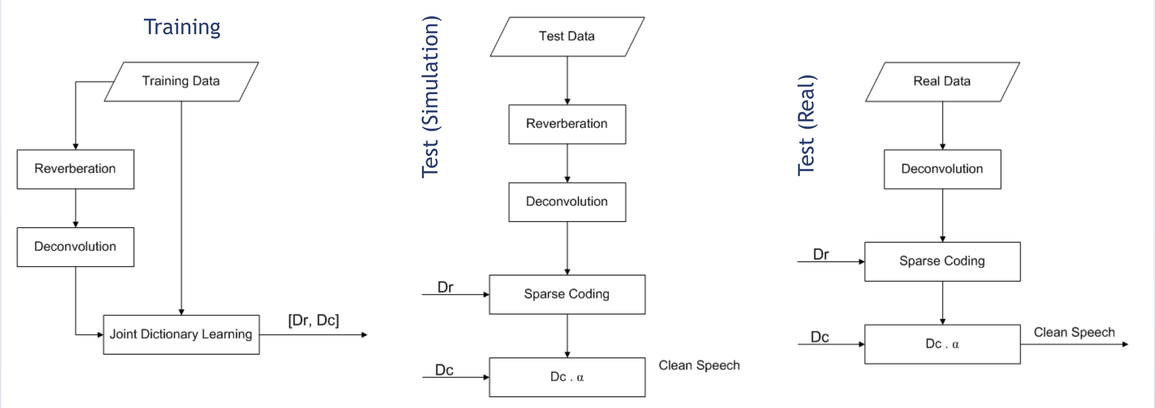
Given $\hat{S}_c$, a speech dictionary D_c and a reverberant dictionary D_r we find spars decomposition of estimated speech in $\underline{D_r}$ using LARC sparse coding algorithm [14]:

$$\hat{S}_c \cong D_r \alpha \tag{11}$$

and multiply known dictionary to sparse code a to make final dereverberated or clean speech estimate:

$$S_c \cong D_c \alpha \tag{12}$$





REFERENCES

[1] G. W. Elco, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3, pp. 229-240, 1996.

[2] A. v. Oppenheim, R. Schafer and T. Stockham Jr, "Nonlinear filtering of multiplied and convolved signals," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 3, pp. 437--466, 1968.

[3] D. Bees, M. Blostein and P. Kabal, "Reverberant speech enhancement using cepstral processing," in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 1991.

[4] G. Xu, H. Liu, L. Tong and T. Kailaath, "A least-squares approach to blind channel identification," *Signal Processing, IEEE Transactions on*, vol. 43, no. 12, pp. 2982--2993, 1995.

[5] H. Kameoka, T. Nakatani and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009.

[6] K. Kumar, R. Singh, B. Raj and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.

[7] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[8] F. Couzinie-Devy, J. Mairal, F. Bach and J. Ponce, "Dictionary learning for deblurring and digital zoom," *arXiv preprint arXiv:1110.0957*, 2011.

[9] M. Aharon, M. Elad and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311--4322, 2006.

[10] M. G. Jafari, M. D. Plumbley and M. E. Davies, "Speech separation using an adaptive sparse dictionary algorithm," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, 2008.

[11] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 5, pp. 1025--1031, 2011.

[12] C. D. Sigg, T. Dikk and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.

[13] D. Barchiesi and M. D. Plumbley, "Dictionary learning of convolved signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.

[14] C. D. Sigg, T. Dikk and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1698--1712, 2012.

[15] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," Springer, 2004, pp. 494--499.

[16] J. Yang, J. Wright, T. S. Huang and Y. Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861--2873, 2010.

[17] D. D. Lee and H. S. Seung, "Algorithms for Non-negative Matrix Factorization," *Advances in Neural Information Processing Systems*, 2001.

[18] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "Wsjcam0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition," in *ICASSP 95*, 1995.

[19] M. Lincoln, I. McCowan, J. Vepa and H. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.