A SPEECH DEREVERBERATION METHOD USING ADAPTIVE SPARSE DICTIONARY LEARNING

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

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

ABSTRACT

We present a monaural blind de-reverberation method based on sparse coding of de-convolved 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 de-convolution 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 de-convolution and sparse coding, and achieves some improvements on objective voice quality testing's like perceptual evaluation of speech quality.

Index Terms—speech dereverberation, sparse coding, dictionary learning, Non-negative Matrix Factor Deconvolution

1 INTRODUCTION

Dereverberation is a highly appropriate and difficult task. Its importance is due to the variety of practical applications while the difficulty arises from the dynamic and long time effect that reverberant environments influence on speech signals. The aim of adaptive dictionary learning is to remove long and non-stationary facts which significantly reduce both speech quality and intelligibility.

Dereverberating speech which is degraded by convolutional noise of reverberant environments (e.g. classrooms, halls, inside car, etc) is both a highly quality improving and difficult task. Its importance is due to its contribution in quality of service in various practical signal processing applications including mobile communications, speech recording and speech recognition. The difficulties are from the dynamic and long time degradation that reverberation affects on speech signals.

Technically speaking, reverberation is the effect of the acoustic channel from the speech source up to the hearing system. The effect of reverberation begins with the production of sound at a location inside a room. The acoustic wave expands radial, reaching walls and other surfaces where energy is both absorbed and reflected. This effect can mainly be modeled by a linear time invariant system:

$$s_r[n] = s_c[n] * h[n] \qquad (1)$$

where, * denotes discrete time linear convolution, $s_c[n]$ is source or clean speech signal, $h[n]$ is the impulse response of a linear system or RIR (Room Impulse Response), $s_r[n]$ is the reverberated signal and n is the time index. The parameters of the filter $h[n]$ change with changes in the environmental parameters such as size of the room, room configuration, position of objects etc. It is typically assumed that the room-response spectral variations rate is slow comparing to the spectral variation rate of speech. As a result, for a short duration (e.g. two or three seconds), we can assume that $h[n]$ is time invariant and thus the entire system in (1) becomes a linear time invariant (LTI) system.

The impulse response of LTI system, $h[n]$, is distinguished into two portions: early reverberation, composed by strong sparse reflections, and late reverberation, characterized by uniform diffusion of the reflections. Ignoring the second part, the impulse response of a room is simplified as a sparse FIR filter which means that the number of reflections or nonzero coefficients of h is small in comparison to the length of the whole RIR which is usually about 100ms.

1.1 Related Work in Derverberation

Blind dereverberation is estimation of the original clean signal from received signal without knowledge of the RIR. Several proposed approaches are classified into two categories by considering whether the inverse RIR is necessary to estimate or not. In fact, all algorithms attempt to obtain clean signal by attenuating the RIR effects or by cancelling it. Simply, a group of algorithms try to lighten the symptoms of the signal degradation, while the other attempt to find a solution for its cause. Regardless of many different algorithms, practical dereverberation is still an open problem.

1) Beamforming: Beamforming is a filtering method designed to configure multiple electrical sensors e.g. microphones to convey a special directivity pattern [1]. This can be used to differentiate a source in a noisy environment or to attenuate the interference caused by reverberation. Speech enhancement of a signal recorded from a far field acquisition is a typical application. This technique can also be used to obtain a higher intelligibility of the diffused signal by designing a loudspeaker array that can focus the acoustic energy in a confined spatial region, minimizing the reflections due to the surrounding walls and objects.

2) Speech Enhancement: Speech dereverberation using enhancement algorithms e.g. peak picking algorithm to identify and remove peaks in cepstrum of speech signal has initially suggested by Oppenheim and Schafer [2] and followed by others e.g. Bees *et al.* [3] later.

3) Blind Deconvolution: Blind system identification or deconvolution is based on the idea that reverberation effect can be modeled as LTI system assuming the possibility to identify multichannel FIR systems based only on the channel outputs under certain conditions [4].
Blind multichannel system identification is usually made by the cross-relation between two records s_{r1} and s_{r2} and the corresponding two slightly different RIRs h_1 and h_2, where the time index is temporarily omitted for simplicity. The cross-relation is given by:

$$s_{r_1} * h_2 = (s_c[n] * h_2) * h_1 = s_{r_2} * h_1 \qquad (2)$$

which leads to the system of equations:

$$\mathrm{R} \times \mathrm{h} = 0 \qquad (3)$$

these equations can be solved using least squares method assuming that: a) RIRs does not include any zeros, b) correlation matrix of clean signal is full ranked [4].

Complementary, in continue to the conventional methods for channel identification, some novel methods are proposed by Kameoka *et al.* [5] and Kumar *et al.* [6] that use monaural signals instead of multiple observations with the aid of sparse and non-negative nature of speech signals in short time Fourier transform or STFT domain.

1.2 Sparse Coding

Sparse coding is the process of finding a vector $x \in \mathcal{R}^n$ from a given signal $y \in \mathcal{R}^m$ and overcomplete ($m \ll n$) matrix $D \in \mathcal{R}^{m \times n}$ known as dictionary such that:

$$\min_x \|x\|_p \quad subject\,to \quad y = Dx \qquad (4)$$

or

$$\min_x \|x\|_p \quad subject\,to \quad \|y - Dx\|_2 \leq \varepsilon \qquad (5)$$

where $\| \ \|_p$ is l^p-norm. This is an optimization problem which obviously does not have a unique solution. Therefore, the process of finding sparsest representation (4) or approximation (5) is typically done by one of the pursuit algorithms. Parameter p depends on coding algorithm e.g. $p=0$ for OMP algorithm and $p=1$ for Lasso or BP algorithm.

Obviously, sparse coding not only depends on the algorithm which finds sparsest code, x, but also depends on dictionary, D, which is required to be as complete as possible. Dictionary matrix D is typically learned by one of the known dictionary learning methods e.g. NMF. Fortunately, dictionary learning is being popular in recent years and various algorithms have proposed for different applications e.g. sparse coding [7], image enhancement [8] and noise cancellation [9], speech source separation [10], speech coding [11], speech enhancement [12] and convolved signals [13].

Despite of algorithms that use l^1–norm for sparse coding, a new method named K-SVD has used l^0-norm instead [9]. Experiments on GRID speech corpus shows that it gives very good results for speech enhancement [14].

1.3 Contributions of Paper

We propose an extension to the novel methods [5] [6] that can be interpreted as an essential non-negative matrix factor deconvolution (NMFD) algorithm [15] using K-SVD dictionary learning [9] to enhance estimated speech signal after deconvolution process by means of special sparse coding algorithm named least angle regression with a coherence criterion or LARC [12].

Adaptive sparse dictionary learning consists of two steps aiming to improve both speech quality and intelligibility: first non-negative blind deconvolution to remove long time influence of room impulse response on the speech signal, second an adaptive sparse coding to compensate variations of recording sight (e.g. microphone movement, changes in speaker position or any reflecting obstacle inside the acoustic environment.

We present a monaural blind dereverberation method based on sparse coding of NMFD version of reverberated speech signal in a dictionary learned by joint dictionary learning method [16], consisting of the concatenation of a clean speech and a deconvolved version of reverberated copy. Suppose we have three different monaural records of the same voice played back inside a small, medium and large room respectively. What we expect from the NMFD algorithm is something like:

$$\begin{aligned} V_{r1} &= V_c *^{\rightarrow} H_1 \\ V_{r2} &= V_c *^{\rightarrow} H_2 \\ V_{r3} &= V_c *^{\rightarrow} H_3 \end{aligned} \qquad (6)$$

where V_{r1}, V_{r2}, V_{r2} are captured speech in small, medium and large room respectively; V_c is clean speech, H_1, H_2, H_3 are their RIRs; and $*^{\rightarrow}$ is defined as following:

$$A *^{\rightarrow} B = \sum_n A^{\rightarrow n} B_n \qquad (7)$$

where the operator $(\ )^{m\rightarrow}$ shifts the columns of its argument by m spots to the right; and B_n is a diagonal matrix made from n[th] column of B.

In practice, we do not reach such a straightforward answer (Fig. 1). Why? Because, to be able to use NMF, we used magnitude spectrogram, which is non-negative, instead of complete STFT; and phase effect is ignored though it contains important information. Therefore, we rewrite (7) as following:

$$\begin{aligned} V_{r1} &= V_{c1} *^{\rightarrow} H_1 \\ V_{r2} &= V_{c2} *^{\rightarrow} H_2 \\ V_{r3} &= V_{c3} *^{\rightarrow} H_3 \end{aligned} \tag{8}$$

Although, V_{r1}, V_{r2}, V_{r2} are different non-negative matrices, they carry *similar information* or speech features. At this stage, we need a block that can translate deconvolved speech features into clean speech atoms. To be able to do this transform, we use joint dictionary learning [16] which can be done using one of the following datasets:

- V_{c1} and proper clean speech
- V_{c2} and proper clean speech
- V_{c3} and proper clean speech

Environment specific dictionary is originally learned off-line on a training corpus for different locations, while adaptive dictionary learning continues on-line for any alternate surroundings. Our approach uses both non-negative blind deconvolution and sparse coding, thus achieves significant improvements on objective voice quality testing's like PESQ.

Theoretically, a sample reverberant environment will be used to simulate behavior of phase which is a random variable. It is not any specific input or constraint to the proposed method. Analytically, there should not be much difference between different dictionaries learned by any of the named datasets. However, in statistical point of view, it's preferred to use concatenation of V_{c1}, V_{c2}, V_{c2} matrices and proper clean speech to learn a general dictionary.

2 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 [14]. 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.

2.1 Motivations

Convolutive reverberation can be expressed in terms of RIR (1). Equivalently, in STFT domain we may write:

$$S_r[n] = \sum_m S_c[m]H[n-m] \tag{9}$$

where Sr, Sc and H are magnitudes of reverberated signal, clean signal and RIR in frequency domain respectively. Equation (9) can also be written in the form of operators as following:

$$S_r \cong \hat{S}_r = \sum_{m=1}^{M} \hat{S}_{cm} H^{m\rightarrow} \tag{10}$$

Non-negative Matrix Factor Deconvolution (NMFD): beginning with the decomposition of non-negative matrix $V \in \mathcal{R}_+^{M\times N}$ into multiplication of two non-negative matrices $W \in \mathcal{R}_+^{M\times P}$ and $H \in \mathcal{R}_+^{P\times N}$ 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]:

$$D = \left\| V \otimes ln\left(\frac{V}{W.H}\right) - V + W.H \right\|_F \tag{11}$$

which yields to an iterative solution:

$$\begin{cases} H = H \otimes \dfrac{W^T \dfrac{V}{W.H}}{W^T.1} \\ W = W \otimes \dfrac{\dfrac{V}{W.H} H^T}{1\ H^T} \end{cases} \tag{12}$$

where $\otimes$ operator is Hadamard product (an element-wise multiplication) and divisions are element-wise too. Setting

$$\Lambda = \sum_{t=0}^{T-1} W_t H^{t\rightarrow} \tag{13}$$

and using cost function

$$D = \left\| V \otimes ln\left(\frac{V}{\Lambda}\right) - V + \Lambda \right\|_F \tag{14}$$

results:

$$\begin{cases} H = H \otimes \dfrac{W_t^T \left(\dfrac{V}{\Lambda}\right)^{\leftarrow t}}{W_t^T \cdot 1} \\ W_t = W_t \otimes \dfrac{\dfrac{V}{\Lambda} H^{t\rightarrow T}}{1 \cdot H^{t\rightarrow T}} \end{cases} \tag{15}$$

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

$$\begin{cases} H = H \otimes \dfrac{\hat{S}_{cm}^T \cdot \left(\dfrac{S_r}{\hat{S}_r}\right)^{m\rightarrow}}{\hat{S}_{cm}^T \cdot 1} \\ \hat{S}_{cm} = \hat{S}_{cm} \otimes \dfrac{\dfrac{S_r}{\hat{S}_r} (H^{m\rightarrow})^T}{1 \cdot H^{m\rightarrow}} \end{cases} \tag{16}$$

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. α; then D is split into two distinct dictionaries D_b and D_s to represent blurry $D_b\alpha$ and sharp samples $D_s\alpha$ in consequence. At test time, given a new blurry sample x, a sparse code α is obtained by decomposing x using D_b, and ones hopes $D_s\alpha$ 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 $\hat{S}_c$ and then use it as a training set similar to what is done in deblurring application as blurry patches. Therefore, we may write:

$$S_j \cong D_j\alpha \tag{17}$$

or

$$\begin{bmatrix} S_c \\ \hat{S}_c \end{bmatrix} \cong \begin{bmatrix} D_c \\ D_r \end{bmatrix} \alpha \tag{18}$$

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

2.2 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.

As discussed in the introduction, we suppose that the observed reverberated speech magnitude is the convolution result of clean speech magnitude S_c and RIR. The goal of the deconvolution step is to obtain an estimate $\hat{S}_c$ 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 $D_c \in \mathcal{R}^{D\times L}$ and a reverberant dictionary $D_r \in \mathcal{R}^{D\times L}$ we find spars decomposition of estimated speech in D_r using LARC sparse coding algorithm [14]:

$$\hat{S}_c \cong D_r\alpha \tag{19}$$

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

$$S_c \cong D_c\alpha \tag{20}$$

2.3 Post-Processing

Although the dereverberation process described so far can obtain reasonable results on objective testing's measured for output sounds, it does not need to be the end of the dereverberation process. We can use some post-processing on the phase of output signal to boost the quality of the results even more. In this section we briefly describe a possible approach using phase continuity of voiced speech signal.

A voiced speech frame can be written as a weighted sum of limited number of sinusoids e.g. N, leading to the harmonic signal model:

$$s[n] = \sum_{n=0}^{N-1} a_n \cos(n\omega_n + \varphi_n) \tag{21}$$

where a_n, φ_n and ω_n are amplitude, phase and normalized angular frequency:

$$\omega_n = 2\pi f_n/f_s = 2\pi(n+1)\, f_0/f_s \tag{22}$$

where, f_s, f_0, and f_n are sampling, fundamental and harmonic frequencies in consequence. The instantaneous phase of each sinusoid term of (18) in continuous time can be expressed as:

$$\varphi(t) = \int^{t} \omega(\tau)d\tau \tag{23}$$

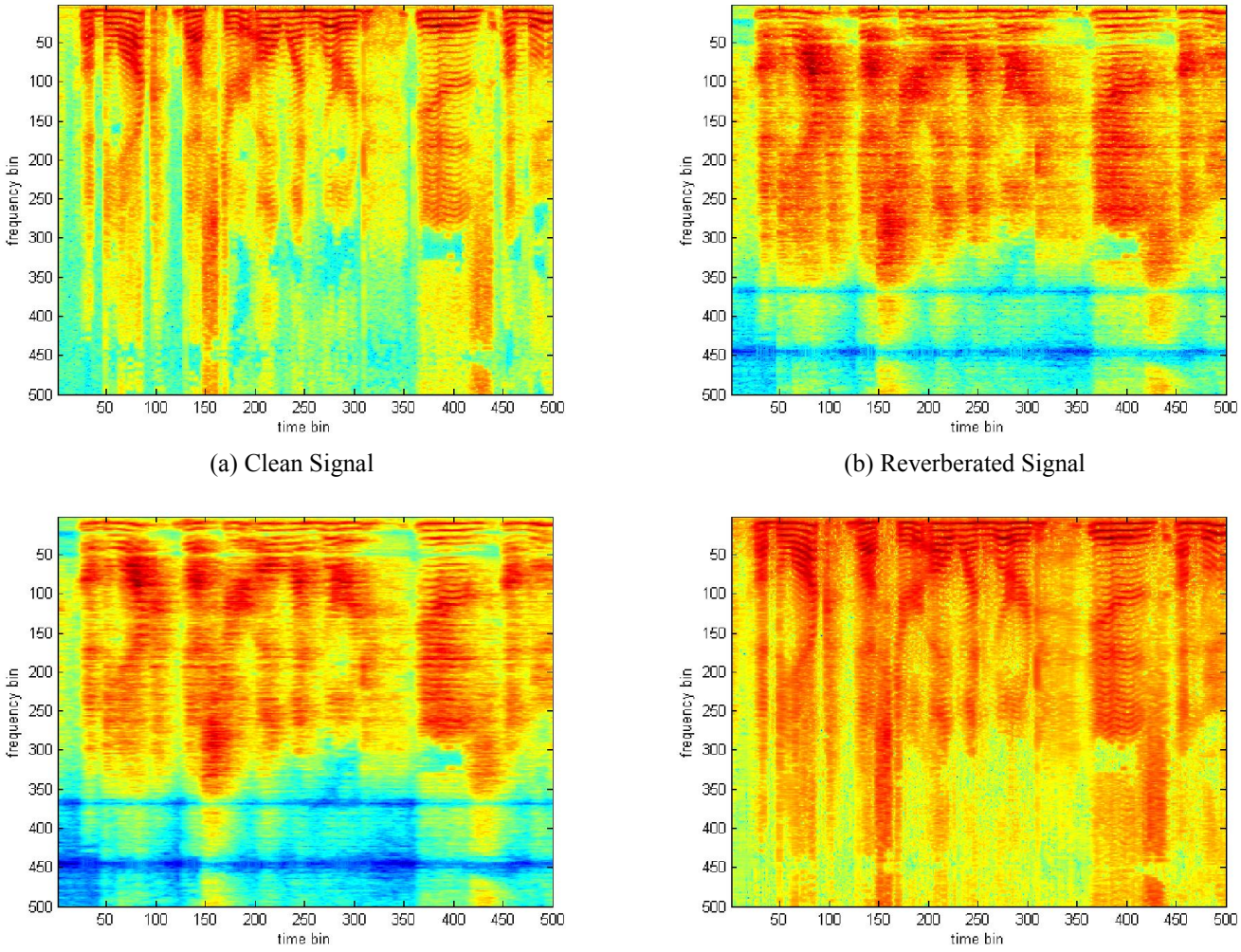
we may assume ω_n constant if the integration interval is [t_{n-1}, t_n], therefore:

$$\begin{aligned} \Delta\varphi_n \cong \int_{t_{n-1}}^{t_n} \omega_n d\tau &= \omega_n(t_n - t_{n-1}) \\ &= 2\pi\frac{f_n}{f_s}N \end{aligned} \tag{24}$$

hence, the phase differences of consecutive analysis of voiced frames n and n-1 in STFT domain for frequency bin m can be estimated as following:

$$\begin{aligned} \Delta\varphi[m,n] &= \varphi[m,n] - \varphi[m,n-1] \\ &= 2\pi\frac{f_n - f_n^{(n)}}{f_s}N \end{aligned} \tag{25}$$

where N is FFT window length and $f_n^{(n)}$ is normalized harmonic frequency of frame n.



(a) Clean Signal

(b) Reverberated Signal

(c) Deconvolved Signal

(d) Dereverberated Signal

Fig. 1. Spectrogram of Clean, Rreverberated, Deconvolved and Dereverberated using joint dictionary learning signals.

Beginning with the initial estimation of clean speech phase of voiced sound, we can compute approximate clean phase of subsequent frames using harmonic frequency of each frame which is directly related to the fundamental frequency or pitch of speech signal.

3 EXPERIMENTS

To evaluate the proposed method, we used the data provided by REVERB2014 which consists of a training set and two test sets; one for development and the other for evaluation. Both development and evaluation test set consist of two different parts, namely simulated data, SimData: utterances from the WSJCAM0 corpus [19]; and real recordings, RealData: utterances from the MC-WSJ-AV corpus [18].

Training utterance of each speaker transformed to STFT domain using discrete time FFT, Hanning window of size 64ms or 1024 samples and 50% overlap or 32ms step size. Reverberated samples of each speaker are also transformed to the STFT domain using the same parameters and deconvolved using NMFD algorithm with 10 iterations. Deconvolved version of training samples concatenated to their respective clean samples to train joint dictionary. Joint dictionary, then split into two distinct dictionaries named *clean dictionary* and *reverberation dictionary*.

Spectrograms of the clean speech signal and the synthesized reverberant signal of a sample speech are illustrated in Fig. 1 (a) and (b) respectively. The spectrograms of the deconvolved signal obtained after NMFD step and final dereverberated signal after applying speech enhancement using joint dictionary learning are shown in Fig. 1 (c), (d).

We used MATLAB™ scripts from REVERB Challenge 2014 which contains different objective testing's except PESQ to measure the dereverberation performance. For the PESQ objective test, we used C source code from ITU

official homepage and compiled it using Visual Studio 2010™ to make a standalone executable program for 64-bit Windows™ operating system.

For the sample case, the present method improved the PEQS from 2.326 to 2.886, while the recently proposed approach using NMFD was only able to improve it to 2.442 (table I).

Overall results show that the proposed method improves all of the objective test values as cepstrum distance, log likelihhod ratio, frequency weighted segmental SNR, SRMR and PESQ.

TABLE I. SAMPLE TEST RESULT

Signal	PESQ	
	Rev. Phase	*Clean Phase*
Clean	-	4.500
Reverberated	2.326	-
Deconvolved	2.442	2.907
Dereverberated	2.886	3.477

Outputs of applying this method on evaluation test set (full batch, monaural) are summarized in tables II to VI.

4 CONCLUSION

In this paper, we have introduced the possibility of using convolutive non-negative matrix factorization in consequence to the sparse dictionary learning to address the reverberation problem. We believe, this kind of enhancement which currently is unusual, will be the primary material of an optimized algorithm for speech signal dereverberation in future.

TABLE II. CEPSTRUM DISTANCE

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	2.67	3.38	2.30	3.05
Far-2	5.21	3.78	5.04	3.43
Far-3	4.96	3.62	4.73	3.34
Near-1	1.99	3.35	1.68	3.02
Near-2	4.63	3.62	4.24	3.25
Near-3	4.38	3.48	4.04	3.10
Average	3.97	3.53	3.69	3.21

TABLE III. SRMR

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	4.58	5.19	-	-
Far-2	2.97	5.03	-	-
Far-3	2.73	4.97	-	-
Near-1	4.50	5.25	-	-
Near-2	3.74	5.23	-	-
Near-3	3.57	5.20	-	-
Average	3.68	5.15	-	-

TABLE IV. LOG LIKELIHOOD RATIO

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	0.38	0.40	0.35	0.35
Far-2	0.75	0.47	0.63	0.41
Far-3	0.84	0.49	0.76	0.43
Near-1	0.35	0.39	0.33	0.34
Near-2	0.49	0.42	0.40	0.36
Near-3	0.65	0.43	0.59	0.38
Average	0.58	0.43	0.51	0.38

TABLE V. FREQUENCY WEIGHTED SEGMENTAL SNR

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	6.68	8.83	9.24	10.98
Far-2	1.04	7.23	1.77	9.65
Far-3	0.24	6.97	0.89	8.97
Near-1	8.12	9.22	10.72	11.38
Near-2	3.35	8.62	5.52	11.26
Near-3	2.27	8.33	4.21	10.61
Average	3.62	8.20	5.39	10.48

TABLE VI. PESQ

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	2.59	2.75	-	-
Far-2	1.99	2.66	-	-
Far-3	1.87	2.62	-	-
Near-1	3.11	2.78	-	-
Near-2	2.39	2.77	-	-
Near-3	2.27	2.76	-	-
Average	2.37	2.72	-	-

TABLE VII. SRMR (REALDATA)

	Mean		Median	
Room	*Org.*	*Enh.*	*Org.*	*Enh.*
Far-1	3.52	4.98	-	-
Near-1	4.06	5.04	-	-
Average	3.79	5.02	-	-

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.