Speech Enhancement Using Constrained Low Rank and Sparse Matrix Decomposition

Joshua Finer
Electrical Engineering
Columbia University, USA
Email: jf2904@columbia.edu

Sarat Chandra Vysyaraju
Electrical Engineering
Columbia University, USA
Email: scv2114@columbia.edu

***Abstract*—Speech enhancement techniques often use strong models for both noise and speech by decomposing the mixture under certain assumptions. Robust Principal Component Analysis (RPCA) and its variants have been used previously in signal separation tasks to separate a low rank background source from a sparse and more variable foreground source. Assuming noise has limited spectral variability or high correlation across time frames, and speech signal being sparse in the time-frequency domain we can apply this approach to speech enhancement. We use a Constrained Low-rank and Sparse Matrix Decomposition (CLSMD) technique which implements an alternative projection algorithm to separate speech from noise. The advantage of this approach is that one does not need to know the exact distribution of the noise signal.**

***Keywords*—Speech Enhancement, RPCA, Low-rank matrix approximation, sparse representation**

I. Introduction

Speech signals in real world environment are often corrupted by various adverse noises such as vehicle noise, channel noise, competing speakers, other environmental noises, etc. Thus, there arises a need to enhance speech quality and intelligibility by performing some sense of noise suppression. This problem has been of great interest over the last few decades and a number of algorithms or techniques have been proposed to address it. The most popular approaches include Spectral Subtraction, Minimum Mean Square Error (MMSE) estimation, Wiener Filtering (WF) and Subspace Methods. However, most of these methods seem to suffer from poor performance under low SNR conditions.

In this report, we discuss about an approach based on the principle of constrained low-rank and sparse matrix decomposition (CLSMD) to address the problem. The idea is inspired by the robust principal component analysis algorithm (RPCA) theory, addressed during the course [1] which states that if an observation matrix is the superposition of a low-rank component and a sparse component, then the two components can be perfectly recovered with a great probability under mild conditions. We utilise this property and model the noisy speech signal in a similar manner as explained further. The speech signal in the time-frequency (T-F) domain is sparse in nature and the noise component in each time-frame is approximately correlated across time over the frames and thus the noise spectrogram can be assumed to be in a low-rank subspace. Thus, we can exploit the approach similar to that suggested by RPCA to reconstruct a clean speech signal from a noisy speech signal.

We initially approached the problem in a manner similar to the that explained in the Singing-Voice Separation [2] paper using the Augmented Lagrange Method of RPCA. However, we observed that the separation of the low-rank noise component and the sparse speech is highly sensitive to the tuning parameter λ which establishes a single relative constraint on both the sparsity and low-rank nature of the components. In-order to ensure better separation, we believe that an individual constraint on the sparsity and the low-rank components should be established as proposed by the CLSMD approach.

The CLSMD method implemented in this project is also favorable due to the following reasons: (1) The method is a non-parametric as there were no specific assumptions made about the distribution of the spectral components of either speech or noise. It only requires that noise is low-rank and speech is sparse in the timefrequency space. (2) Unlike many traditional speech enhancement methods, the algorithm does not require Voice Activity Detection (VAD) methods to estimate the noise spectrum, as the speech and the noise components are recovered simultaneously using the alternating projection algorithm.

The rest of the report is organised as follows. Section 2 explains the RPCA method and its implementation for the speech enhancement problem along with the difficulties observed in the approach. Section 3 describes the CLSMD method and its optimization algorithm followed by the CLSMD based speech enhancement system. This is followed by the Experimental Results and their Interpretation in Section 4. In Section 5, we briefly discuss the inferences and the key learnings of the project.

II. RPCA Approach

The basic RPCA method aims to solve the following problem. Suppose we are given a data matrix Y which is assumed to be a superposition of a low-rank matrix L and a sparse matrix S. Theoretical results show that under certain rather weak assumptions we can recover the low rank and sparse component by solving

$$\min_{L,S} ||L||_* + \lambda ||S||_1 \quad s.t.\ Y = L + S \tag{1}$$

The nuclear norm serves as a surrogate for the rank of L and the l^1 norm serves as a surrogate for the sparsity of S. $\lambda > 0$ is a parameter that trades off the sparsity of S

with the low-rankness of L in the sense that as λ increases, the decomposition of Y has a fewer non-zero entries for S and higher rank for L. The fine tuning of this parameter is a significant factor in the quality of the decomposition with respect to the problem at hand.

The equality constraint for RPCA can be relaxed to handle small random perturbations in the matrix Y. This yields the problem of Stable RPCA presnted as follows.

$$\min_{L,S} \frac{1}{2}||Y - L - S||_F^2 + \lambda_1||L||_* + \lambda_2||S||_1 \quad (2)$$

Generally while applying RPCA, the goal is to recover L. However, in the context of the problem we are studying, Y represents the magnitude spectrum of the short-time Fourier transform of our speech signals. Here, we wish to recover S which should correspond to the magnitude spectrum of the denoised speech and so L is not the object of interest to recover. Both these problems, the stable and non-stable versions of RPCA, are convex and as such they have unique global optimum. There exist efficient algorithms including Augmented Lagrangian Method (ALM) and Accelerated Proximal Gradient (APG) for solving them [6]. ALM can be used to solve the equality constrained version of RPCA in (1). APG can solve the relaxed version in (2).

We have applied RPCA to several noisy test signals containing speech sentences as described later. The resulting decomposition is shown in Figure 1 for $\lambda = .4, .8$ and 1.2. As we can see the low rank part and sparse components are very sensitive to the lambda parameter. For low values of λ, the rank of L does not capture all the noise from the speech as seen in the top row. As we increase the value of λ to 1.2, the speech gets absorbed into the low rank part L and therefore lost. Since there is a delicate balance between the sparsity and low-rankness of the separation, it might be worthwhile to impose hard constraints on L and S so as to maximize separation. This leads to the CLSMD formulation for this problem. It is worthwhile to note that the APG method was used for solving (2) as well. However, this yielded similar decompositions to ALM where the rank of L was nearly identical to the rank of L using APG.

III. CLSMD Approach

The CLSMD method uses a formulation of the problem as follows.

$$\min_{L,S} ||Y - L - S||_F^2 \quad (3)$$
$$s.t.\ \ rank(L) \leq r, ||S||_1 < h, S_{i,j} \geq 0$$

Here, r is an upper bound on the rank of L, and h bounds the sparsity of S. This algorithm can be viewed as a supervised low-rank and sparse decomposition method [3]. A natural approach described in the paper to optimizing this objective with non-convex constraints is to use an alternating projection optimization algorithm by solving the following two subproblems.

$$L_t = \arg\min_{L,S} ||Y - L - S_{t-1}||_F^2 \quad (4)$$
$$S_t = \arg\min_{L,S} ||Y - L_t - S||_F^2$$

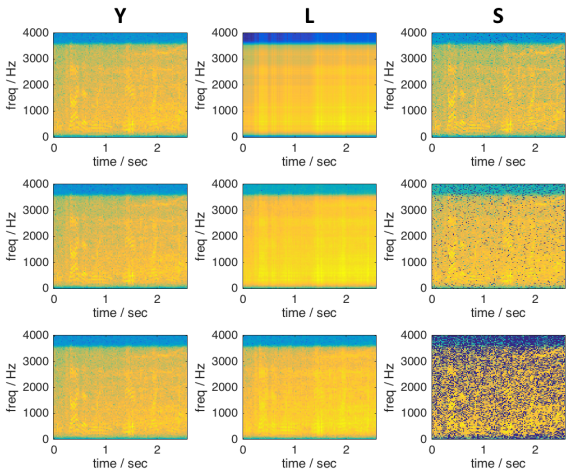

Fig. 1: RPCA separation of Y into L and S for λ = 0.4, top row, 0.8, middle row and 1.2, bottom row. The corresponding ranks for each L are 1, 46, and 114.

When updating L_t we truncate the SVD of $Y - S_{t-1}$ to yield a rank r approximation of L. Then, when updating S_t we use a hard-thresholding operator on $T - L_t$ with parameter T, corresponding to the sparsity constraint on S. A larger value of T enforces a higher degree of sparsity on S at each iteration. Pseudocode for the algorithm is given in Figure 2 as described in the paper.

Algorithm 1: Optimization algorithm for CLSMD

Given r, T, ε, t_{maxiter};
Initialize $Y_0 = M$, $S_t = [0]_{N\times K}$, t=0;
  while not converged do
    % Update of low-rank matrix L
    $U\Lambda V^T = svd(Y_t)$;
    $L_t = \sum_{i=0}^{r} \lambda_i U_i V_i^T$;
    % Update of sparse matrix S
    $X_t = Y_t - L_t + S_t$;
    $S_t = X_t \otimes (X_t > T)$;
    % Stopping criteria
    If $||M - L_t - S_t||_F^2 / ||M||_F^2 \leqslant \varepsilon$ or $t == t_{\text{maxiter}}$
      break;
  end
    $Y_t = L_t + X_t - S_t$
    $t = t + 1$;
  end while
  output: $L = L_t$, $S = S_t$

Fig. 2: CLSMD Algorithm

A. Problem Formulation

We are given a set of noisy speech signals obtained from the Noizeus database [3]. The set contains 30 sentences from male and female speakers corrupted by different types of noise: suburban train noise, babble, car, exhibition hall, restaurant,

street, airport and train-station at various SNRs: 0-15 dB. All files were sampled at 8 kHz sampling rate. The goal is to apply this algorithm and modify the signal in the time-frequency domain to obtain a deionised version of the original signal. The performance of the system is measured by segSNR and PESQ of the estimated reconstructed signal. This process is accomplished within the AMS framework which is shown in the diagram in Figure 3

B. Technical Approach

Based on the previous efforts that have applied RPCA successfully to the separation of foreground singing and background music [2], we have attempted to make use of the CLSMD algorithm to separate noise and speech, thereby enhancing the quality and intelligibility of the signal. When we run our algorithm, we comp[ute the spectrogram using a 1024 point FFT and a window of 37.5 ms (300 points) and a hop of 10 ms (80) points. Our technical approach to signal reconstruction follows the Analysis, Modification and Synthesis (AMS) framework for short-time speech processing. The phase of the original noisy signal is kept for multiplying with the estimated magnitude prior to inverting the STFT. The re-synthesis is facilitated using the overlap-and-add approach. The time-frequency masking is accomplished by multiplying the estimated magnitude S by a binary mask according to the formula

$$B_g(i,j) = \begin{cases} 1 & \text{if } |S(i,j)| > g \cdot |L(i,j)| \\ 0 & \text{otherwise} \end{cases}$$

where the gain parameter is manually chosen to be 5.

A depiction of the CLSMD decomposition for a signal corrupted at 0 dB is given in Figure 4. The results show that the decomposition enforces L to have low rank while the sparsity threshold parameter T constrains S to be sparse. Thus, the noise in the speech signal is suppressed, relative to RPCA.

An illustration of the effect of the masking is given in Figure 5 for the same signal. Masking reduces noisy artifacts in the T-F domain by setting any component of $S < g \cdot |L|$ to zero, where $g = 5$.

The reconstruction of a sample noisy signal at SNR of 0dB using CLSMD is given in Figure 6. It is compared to the original clean speech. The enhancement appears to be improved as a result of the masking.

IV. Experiments & Interpretation

In this section we report and analyze a series of experiments to asses the performance of the CLSMD decomposition as we vary the sparsity parameter T and the rank parameter r for the data set of corrupted speech signals at various types and levels of noise. We also explore the effect of binary masking.

A. Experimental settings

A total of 30 speech sentences (sp1 - sp30) have been used from the NOIZEUS database [5] to analyse the speech enhancement technique using the CLSMD method, as explained in the previous section. The sentences were corrupted by 7 types of noises: car, exhibition, street, airport, station, babble and restaurant at 0, 5, 10 and 15dB. Both clean and noisy

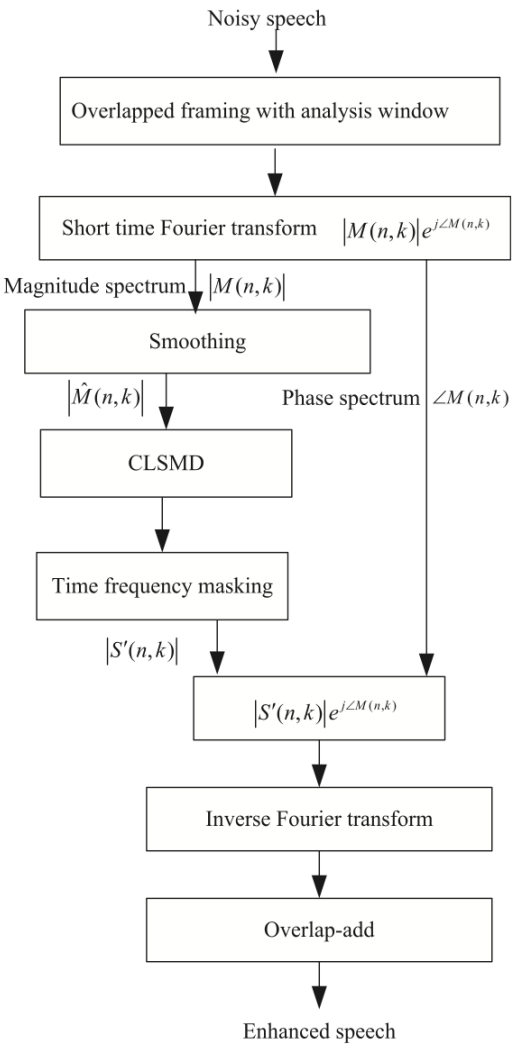

Fig. 3: AMS Framework

speech signals were sampled at 8kHz 16 bits. These signals were segmented into frames of 37.5 ms (300 samples) with 40% overlap and then transformed into time-frequency domain by performing a 1024 point STFT.

B. Evaluation Measures

In-order to quantify and analyse the performance of the method implemented, we use two objective performance measures : (1) Segmental SNR (segSNR) and (2) Perceptual Evaluation of Speech Quality (PESQ). The segSNR measure primarily quantifies the effectiveness of noise suppression in the speech signal and is defined as

$$segSNR = \frac{1}{|P|} \sum_{p \in P} 10 \log_{10} \left(\frac{||s_p||^2}{||s_p - \hat{s_p}||^2} \right) \qquad (5)$$

where s_p and $\hat{s_p}$ denote the original clean speech & enhanced speech time frame respectively and P is the index set which denotes all clean speech frames. The segSNR thus

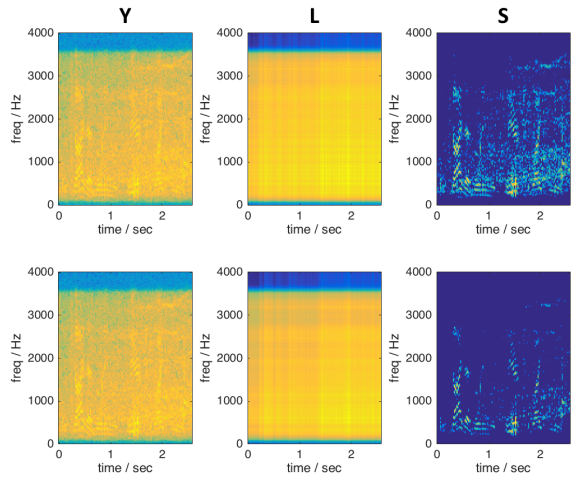

Fig. 4: CLSMD separation of Y into L and S for T = 0.2, top row and T = 0.5, bottom row. The corresponding ranks for each L are both 1

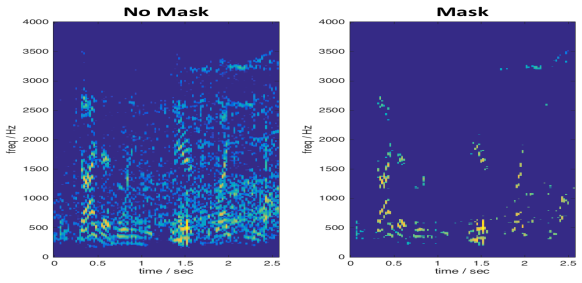

Fig. 5: The Effect of Binary Maksing (g=5)

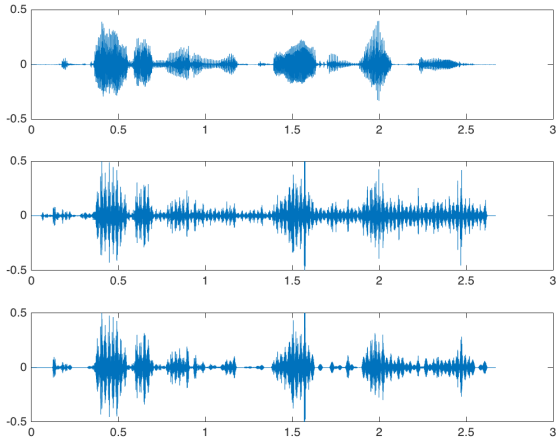

Fig. 6: Clean speech (top), Enhanced Speech without Binary Masking (middle) and Enhanced Speech with Binary Masking (bottom)

calculates the SNR only for frames containing speech. Secondly, the PESQ measure quantifies the overall quality of the reconstructed speech. We used the MATLAB implementations of both the measures shared by the Speech Enhancement paper [4].

C. *Influence of sparsity and low-rank constraints*

In the CLSMD approach, the reconstruction of the sparse (speech) and low-rank (noise) components and their separability from the mixed signal is primarily governed by the sparsity and the low-rank constraints. In-order to understand the influence of these parameters, we have conducted the experimental analysis on the entire data set for sparsity constraint (T) varying from 0.1 to 1 in steps of 0.1 and the low-rank constraint (r) being 1 or 2. In the analysis all the segSNR and PESQ measures were averaged over the entire set of 30 sentences for each values of r, T and type of noise.

As discussed in the previous section, we would also like to emphasise the importance of the binary filtering technique and its influence in improving the speech quality in this approach. Thus we calculated the segSNR measures as mentioned above for two cases of speech reconstructed signals, i.e, before and after filtering the signal using binary filter. And it does play an interesting role in altering the effect of sparsity and low-rank constraints. The results indicated that the performance of the segSNR feature of the data under strong noise conditions (low SNR values) had been best for high values of T, as restricting the sparsity of the speech component led to performing better noise suppression, but at the expense of losing valuable speech related information. So there exists a need to establish a trade-off between the intelligibility of speech and the SNR of the signal. However, as we apply the binary filter to the reconstructed speech data, these measures flatten for an optimal value of the sparsity constraint T and thereby establishing the necessary trade-off.

Figure 7 illustrates the influence of the sparsity constraint parameter T on speech signals corrupted by street noise. We can observe from the first subplot indicating the segSNR values for reconstructed signals before applying the binary filter, that for the speech signals of type 0 dB and 5 dB (under strong noise conditions), the segSNR values improve as we increase the value of T. The high values of sparsity constraint performs better noise suppression, but consequently a lot of speech information is lost thereby compromising the speech intelligibility of the reconstructed signal. Now if we observe the segSNR values for the reconstructed speech signals after applying the binary filter as illustrated in the subplot 2 of Figure 7, we can clearly see that the segSNR values flatten for the sparsity constraint values near $T = 0.3$ thus maintaining speech intelligibility and also establishing higher speech quality. Moreover, such behavior was observed for the other types of noise that the segSNR values flattened at nearly same values of T, thus we choose $T = 0.3$ as the optimal value of sparsity constraint parameter for our analysis.

Using the optimal value of the sparsity constraint we illustrate the segSNR and PESQ measures in the Figures 8 for all types of noise signals for the low-rank constraint values of $r = 1, 2$ averaged over all the speech sentences. It has been observed that the performance had been better for the low-rank constraint being r=1. We found that increasing the values of r to be large had a negative impact on the performance. This is

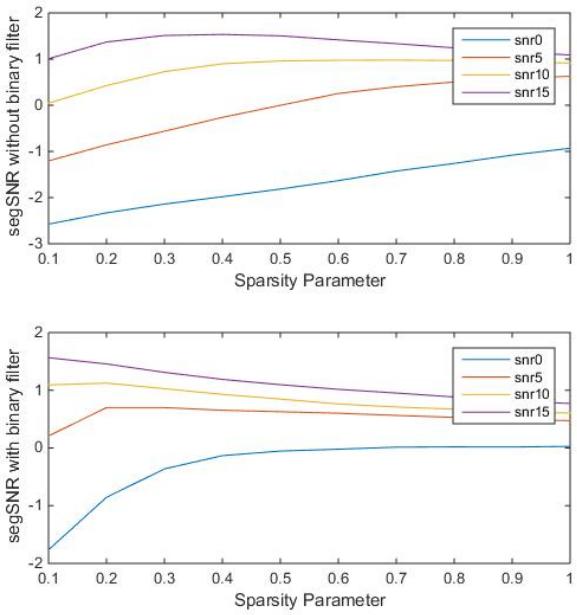

Fig. 7: Influence of T and binary filtering on segSNR parameter for signals corrupted by street noise

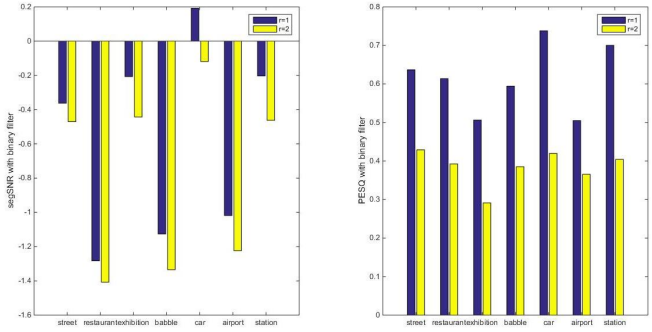

Fig. 8: Influence of r on segSNR and PESQ measures for for all types of noise averaged over all sentences for $T = 0.3$

likely due to the fact that allowing L to have too large a rank will cause it to model too much of the speech information. Therefore, it has to have very low rank.

Thus we observed that the algorithm performs best, as we establish optimal strict constraints on both the sparsity and the low-rank model of data.

V. Inferences

This project gave us great insights on the topic of matrix decompositions. For most of the decompositions much of the speech information was retained in the low frequency bands. Therefore we believe this approach can be integrated with a dictionary model for reconstructing higher frequency bands of speech, thereby improving the quality of the speech sound. Our theoretical understanding of RPCA and its recovery algorithms as well as our understanding of the practical implications has been improved as a result of this work.

Acknowledgment

We would like to thank Prof. John Wright for his beautiful course on Sparse Representation on High Dimension Geometry and his continuous support to help us through out the project.

References

[1] Wright, J., Peng, Y., and Ma, Y., *Robust principal component analysis exact recovery of corrupted low-rank matrices by convex optimization*, NIPS, 2014

[2] Huang, P., Chen, S., Smaragdis, P., and Johnson, M., *Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis*, ICASSP, 2012

[3] C. Sun, Q. Zhu, and M. Wan, *A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition*, Speech Communication, vol. 60, pp. 4455, 2014

[4] Loizou, P.C., *Speech Enhancement: Theory and Practice*, Taylor & Francis, New York, 2009

[5] Hu, Y., and Loizou, P., *Subjective evaluation and comparison of speech enhancement algorithms*, Speech Communication, 49, 588-601. 2007

[6] Z. Lin, M. Chen, L. Wu, and Y. Ma, *The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices*, (UIUC Technical Report UILU-ENG-09-2215, November 2009).

[7] Z. Chen and D.P.W. Ellis, *Speech enhancement by sparse, low-rank and dictionary spectrogram decomposition*, in WASPAA, 2013.