Noise Reduction Based on Robust Principal Component Analysis

Chengli SUN\textsuperscript{1,*}, Qin ZHANG\textsuperscript{2,}, Jian WANG\textsuperscript{2,}, Jianxiao XIE\textsuperscript{2}

\textsuperscript{1}Key Laboratory of Image Processing and Pattern Recognition, Nanchang Hangkong University, Nanchang 330063, China
\textsuperscript{2}School of Information, Nanchang Hangkong University, Nanchang 330063, China

Abstract

In this paper, we present a new speech enhancement method based on robust principal component analysis. In the proposed method, noisy signal is transformed into time-frequency domain where background noise is assumed as a low-rank component and human speech is regarded as a sparse compone. An inexact augmented Lagrange multipliers algorithm is conducted for solving the noise and speech separation problem. Experimental results show the RPCA based speech enhancement method can steadily obtain higher noise suppression performance in noisy conditions, compared to many traditional methods.

Keywords: Noise Suppression; Robust Principal Component Analysis; Matrix Separation

1 Introduction

Background noise acoustically added to speech can decrease the performance of digital signal processing used for applications such as speech compression and recognition. The main objective of speech enhancement is to reduce the influence of noise. Most speech enhancement systems are based on the separation of speech and noise in the spectrum domain using the Short-Time Fourier Transform (STFT), such as spectral subtraction (SS) [1], minimum mean square error (MMSE) estimation [2, 3], Wiener filtering (WF) [4-7], etc.

Traditional speech enhancement systems usually employ voice activity detection (VAD) algorithm to estimate the statistics of the noise signal. However, there are two main shortcomings associated with these methods. Firstly, the noise estimation manner in traditional methods is only suitable for stationary background noise, for example white noise. However non-stationary noises, with noise spectrum levels changing in time, cannot be tracked adequately by noise estimation

\textsuperscript{*}This article is partly supported by National Nature Science Committee of China (No. 61362031, 61203243, 61263040, 61263032), Science and Technology on Avionics Integration Laboratory and China's Aviation Science (No. 20115556007), and the Fund of Jiangxi Provincial Department of Education (No. GJJ12415).

\textsuperscript{*}Corresponding author.
\textit{Email address:} sun.chengli@163.com (Chengli SUN).
and updation only during speech pause period [8]. Secondly, because current VAD techniques are still not reliable at very low SNR conditions, the speech enhancement performance may significantly deteriorate in adverse environment. Moreover, even if the VAD is reliable, changes in the noise spectrum occurring during an active speech cannot influence the noise estimate in a timely manner, this would result in poor estimates during long speech segments with few pauses available [9].

Principal Component Analysis (PCA) is arguably the most widely used statistical tools for data analysis and feature exaction. However, its brittleness with respect to grossly corrupted observations often puts its validity in jeopardy. Unfortunately, gross errors are now ubiquitous in modern applications such as image enhancement, web data analysis, and speech signal processing, where some measurements may be arbitrarily corrupted (due to occlusions, malicious tampering, or sensor failures). Recently, Candes et al. [10, 11] proposed a new theory called RPCA, which can remedy the deficiency of PCA. Assume the collected data matrix \( Z \in \mathbb{R}^{N \times K} \) has an underlying low-rank structure, and a fraction of entries have been corrupted by sparse errors or outliers with arbitrarily large magnitude. Denote these two ingredients as \( P,Q \in \mathbb{R}^{N \times K} \), Candes et al. [11] have shown that under rather broad conditions, one can exactly recover the low-rank matrix \( P \) from \( Z \) as long as the matrix \( Q \) is sufficiently sparse (relative to the rank of \( P \)), and this can be achieved by solving the following optimization problem [12]:

\[
\min \|P\|_* + \lambda \|Q\|_1, \quad \text{subject to} \quad Z = P + Q
\]  

Where \( \|Q\|_1 \) denotes the l1-norm of matrix \( Q \), which is defined as the sum of the absolute values of matrix elements. \( \|P\|_* \) is nuclear norm of matrix \( P \) which equals to the sum of the singular values of \( P \). This problem is known to have a stable solution provided \( P \) and \( Q \) are sufficiently incoherent. A variety of convex relaxation methods have been proposed for solving this problem, such as singular value thresholding [12], augmented Lagrange multiplier method (ALM) [13], and accelerated proximal gradient (APG). For illustration, an example of RPCA decomposition is shown in Fig. 1. In which we can see that the RPCA can recover the low-rank and sparse matrix from highly corrupted matrix accurately. In this paper, we employ apply the RPCA theory to speech and noise separation problem and proposed a RPCA based speech enhancement method. Compared with traditional methods, the proposed speech enhancement method does not need a voice activity detector to find pause in speech segments for noise estimation. Experimental
results show it can obtain higher performance in noisy conditions, compared to many traditional methods.

2 RPCA for Noise Reduction

2.1 Modeling noise reduction as a RPCA formulation

In the speech enhancement scene, since background noises with different time frames are usually highly correlated with each other, it can be assumed to lie in a low-rank subspace. While human speeches have more variation and relatively sparse in the spectral domain, it can be assumed have sparse and high rank structure. Therefore, we can separate speech and noise via RPCA theory, where low-rank component corresponds to background noise and sparse component corresponds to speech. In light of above assumption, we propose a RPCA based speech enhancement method.

We consider the problem of denoising of a noisy signal contaminated by an independent additive noise. Let \( q(t) \) and \( p(t) \) denote the clean speech signal and noise signal, respectively. The observed noisy speech signal \( z(t) \) is expressed as

\[
z(t) = p(t) + q(t) \tag{2}
\]

Since speech can be assumed to be quasi-stationary, it is analyzed frame-wise using short-time Fourier analysis (STFT) \[15\]. The STFT of the corrupted speech signal \( z(t) \) is given by

\[
Z(n, k) = \sum_{m=-\infty}^{\infty} z(m)w(n - m)e^{-j2\pi km/L}, k \in \{1, ..., L\} \tag{3}
\]

Where \( k \) refers to the index of the discrete acoustic frequency, \( L \) is the length of frequency analysis, \( n \) is the index of time-frame, and \( w(n) \) is an analysis window function. The noisy speech signal is divided into frames using a Hamming window function, and then the segmented data are transformed into frequency domain by short-time Fourier transform. The following examples demonstrate the usage of the above environments.

In spectrum domain, we smooth the spectrum magnitude \( |Z(n, k)| \) at each frame by averaging the magnitude values among adjacent three frames, and stack every frames of the signal magnitude spectrum as column vectors over time sequences. Assume the phase difference of speech and noise is zero, we obtain a matrix representation form being

\[
Z = P + Q \tag{4}
\]

In Eq. (4), \( Z = [||Z(n, k)||]_{N\times L} \), \( Q = [||Q(n, k)||]_{N\times L} \), \( P = [||P(n, k)||]_{N\times L} \) are time-frequency matrices associated with the \( z(t) \), \( p(t) \), and \( q(t) \), respectively. The low-rank matrix \( P \) and sparse matrix \( Q \) can be estimated by the RPCA optimization algorithm. In this paper, we adopt inexact augmented Lagrange multipliers (IALM) algorithm for estimating \( P \) and \( Q \). The detail of IALM algorithm is shown in the next subsection.

Due to phase spectrum provided no significant perceptual difference to the enhanced signals \[14\], we combine the entries of \( Q \) with the phase of the original noisy signal to produce an estimate of the Fourier transform of \( q(t) \).
\[ \bar{Q}(n, k) = |Q(n, k)|^{1/2} e^{j\angle Z(n, k)} \]  

Finally, the enhanced speech signal \( \bar{q}(t) \), is constructed by taking the inverse Fourier transform of the spectrum \( \bar{Q}(n, k) \) and followed by least-squares overlap-add synthesis [14].

### 2.2 IALM algorithms for speech time-frequency matrix recovery

The exact ALM (EALM) method [13] is proven to have a pleasing Q-linear convergence speed, while the APG is in theory only sub-linear. A slight improvement over the exact ALM leads an inexact ALM (IALM) method, which converges practically as fast as the exact ALM, but the required number of partial singular value decomposition (SVD) is significantly less. Previous experimental results show that IALM is at least five times faster than APG, and its precision is also higher. In particular, the number of non-zeros computed by IALM is much more accurate than that by APG, which often leave many small non-zero terms in \( Q \).

For the RPCA problem (1), we apply the augmented Lagrange multiplier method by identifying:

\[ X = (P, Q), \quad f(X) = \|P\|_s + \lambda \|Q\|_1, \text{ and } \quad h(X) = Z - P - Q. \]  

Then the Lagrangian function is written as:

\[ L(P, Q, Y, \mu) = \|P\|_s + \lambda \|Q\|_1 + \langle Y, Z - P - Q \rangle + \frac{\mu}{2} \|Z - P - Q\|_F^2 \]  

Where \( \mu \) is a positive scalar, and the ALM method for solving the RPCA problem be described with inexact ALM Algorithm. The initialization \( Y_0^* = \text{sgn}(Z)/J(\text{sgn}(Z)) \) in the algorithm is inspired by the dual problem as it is likely to make the objective function value \( \langle Z, Y_0^* \rangle \) reasonably large. Although the objective function of the RPCA problem is non-smooth and hence the results in [15] do not directly apply here, it has the same excellent convergence property. For EALM Algorithm, any accumulation point \( (P^*_k, Q^*_k) \) of \( (P^*_k, Q^*_k) \) is an optimal solution to the RPCA problem and the convergence rate is at least \( O(\mu_k^{-1}) \) in the sense that

\[ \|\|P^*_k\|_s + \lambda \|Q^*_k\|_1 - f^*\| = O(\mu_k^{-1}). \]  

where \( f^* \) is the optimal value of the RPCA problem. We see that if \( \mu_k \) grows geometrically, the EALM method will converge Q-linearly; and if \( \mu_k \) grows faster, the EALM method will also converge faster. However, numerical tests show that for larger \( \mu_k \) the iterative thresholding approach to solve the sub-problem \( (P^*_k, Q^*_k) = \arg \min_{P,Q} L(P, Q, Y_k^*, \mu_k) \) will converge slower. As the singular value decomposition (SVD) accounts for the majority of the computational load, the choice of \( \mu_k \) should be judicious so that the total number of SVD is minimal. Fortunately, we need not have to solve the sub-problem \( (P^*_k, Q^*_k) = \arg \min_{P,Q} L(P, Q, Y_k^*, \mu_k) \) exactly. Rather, we can solve \( L \) and \( S \) by means of IALM algorithm [14], described in Algorithm 1. Although the exact convergence rate of the IALM is difficult to obtain in theory, extensive numerical experiments have shown that it still converges Q-linearly [14].

#### Algorithm 1 (Matrix recovery by means of IALM)

Input: Observation matrix \( Z \) and \( \lambda \).
\[(1)\] \[Y_0 = Z/J(Z); Q_0 = 0; \mu_0 > 0; \rho > 1; k = 0.\]

(2) while not converged do

(3) //lines 4-5 solve \[P_{k+1} = \arg\min_P L(P, Q_k, Y_k, \mu_k).\]

(4) \[(U, S, V) = \text{SVD}(Z - Q_k + \mu_k^{-1}Y_k);\]

(5) \[P_{k+1} = US_{\mu_k^{-1}}[S]VT;\]

(6) //line 7 solve \[Q_{k+1} = \arg\min_Q L(P_{k+1}, Q, Y_k, \mu_k).\]

(7) \[Q_{k+1} = S_{\mu_k^{-1}}[Z - P_{k+1} + \mu_k^{-1}Y_k];\]

(8) \[Y_{k+1} = Y_k + \mu_k(Z - P_{k+1} - Q_{k+1}); \mu_{k+1} = \rho \mu_k.\]

(9) \[k = k + 1.\]

(10) end while

Output: \((P_k, Q_k)\).

3 Experiments

Performance of the proposed speech enhancement method is evaluated with Noisex-92 database and NOIZEUS database. A total of 15 sentences (sp01∼sp15) were used to evaluate the performance of the proposed estimators. The sentences were corrupted by babble, hfchannel, factory and white noise from Noisex-92 and car, train, street and station noise from NOIZEUS at 0, 5, 10 dB. A sampling frequency of 16 kHz was used. The signal was split up into frames of length 200 samples and a window overlap factor of 50%, and 256 points STFT was adopted for transforming the signal into spectral domain. The parameters setting of RPCA method as follows: maximal iterations of thresholding algorithm is 1000, the frame length is 300, and frame shift percentage is 40%. In order to quantify performance of speech enhancement method, we adopt segmental-SNR measures, which is defined as

\[
\text{segSNR} = \frac{1}{|\Omega|} \sum_{t \in \Omega} 10\log_{10} \left( \frac{\|s_t\|^2}{\|s_t - \tilde{s}_t\|^2} \right).
\]

Where \(s_t\) and \(\tilde{s}_t\) denote a clean and enhanced time-domain signal frame, and is an index set which denotes all clean speech frames with energy within 35 dB of the maximum clean speech frame energy. The segSNR is calculated by computing the SNR only for frames where speech is present. The segSNR measure reflects noise suppression ability. Three speech enhancement methods were adopted for comparison, which include spectral subtraction (SS) \[1\], parametric spectral subtraction (PSS), and MMSE_NPS (minimum mean-square error using a noncausal priori SNR) method \[16\].

Tables 1 shows the performance comparisons based on segSNR measure for the noises from Noisex-92 database at different SNR levels. In terms of segSNR, higher value indicates better
Table 1: Performance comparisons for noises from Noisex-92 database, in terms of segSNR

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>0dB</th>
<th>5dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
<td>SS</td>
<td>-1.2812</td>
<td>0.6329</td>
<td>2.6235</td>
</tr>
<tr>
<td>white</td>
<td>PSS</td>
<td>-1.6643</td>
<td>-0.3489</td>
<td>0.9478</td>
</tr>
<tr>
<td>white</td>
<td>MMSE</td>
<td>-0.7462</td>
<td>0.3414</td>
<td>1.5269</td>
</tr>
<tr>
<td>white</td>
<td>RPCA</td>
<td>1.9169</td>
<td>2.4431</td>
<td>2.8955</td>
</tr>
<tr>
<td>hfchannel</td>
<td>SS</td>
<td>-1.1725</td>
<td>0.2847</td>
<td>2.1479</td>
</tr>
<tr>
<td>hfchannel</td>
<td>PSS</td>
<td>-1.7141</td>
<td>-0.6188</td>
<td>0.4408</td>
</tr>
<tr>
<td>hfchannel</td>
<td>MMSE</td>
<td>-0.8598</td>
<td>0.0229</td>
<td>1.2769</td>
</tr>
<tr>
<td>hfchannel</td>
<td>RPCA</td>
<td>1.7859</td>
<td>2.4942</td>
<td>2.9446</td>
</tr>
<tr>
<td>babble</td>
<td>SS</td>
<td>-1.6994</td>
<td>0.2798</td>
<td>1.8652</td>
</tr>
<tr>
<td>babble</td>
<td>PSS</td>
<td>-2.6209</td>
<td>-1.4963</td>
<td>-0.1251</td>
</tr>
<tr>
<td>babble</td>
<td>MMSE</td>
<td>-2.9161</td>
<td>-1.0856</td>
<td>0.3287</td>
</tr>
<tr>
<td>babble</td>
<td>RPCA</td>
<td>-1.5145</td>
<td>-0.1809</td>
<td>0.9705</td>
</tr>
<tr>
<td>factory</td>
<td>SS</td>
<td>-1.2104</td>
<td>0.8171</td>
<td>3.1467</td>
</tr>
<tr>
<td>factory</td>
<td>PSS</td>
<td>-2.1180</td>
<td>-0.5952</td>
<td>0.7244</td>
</tr>
<tr>
<td>factory</td>
<td>MMSE</td>
<td>-1.6198</td>
<td>-0.2231</td>
<td>1.5129</td>
</tr>
<tr>
<td>factory</td>
<td>RPCA</td>
<td>0.8040</td>
<td>1.6064</td>
<td>2.3502</td>
</tr>
</tbody>
</table>

noise reduction performance. As we can see, the proposed RPCA based method yielded the highest segSNR values at 0 dB SNRs. Among all the noise types we can find that the white and hfchannel noises obtained comparatively high segSNR values against other noises.

Tables 2 shows the performance comparisons based on segSNR measures for the noises from NOIZEUS database at different SNR levels. From which we can see the proposed RPCA method yielded the comparatively higher segSNR scores in all the tested noisy speeches. It is found that the car, train and station noises obtain relatively high segSNR values compared with street noises. Tables 1 and 2 indicate that RPCA method have good noise suppression ability in low SNR situation.

4 Conclusions

In this paper, we presented a RPCA based speech enhancement approach. The advantage of this method is that it can directly estimate enhanced speech and do not need voice activity detector for noise estimation. Moreover, it can obtain high noise suppression performance in low SNR levels. In future research work, we will investigate incorporating more prior knowledge about the speech and noise into the RPCA framework, such as the human perception and temporal continuity property, to further improve the accuracy for the proposed method.
Table 2: Performance comparisons for noises from NOIZEUS database, in terms of segSNR

<table>
<thead>
<tr>
<th>Noise</th>
<th>Method</th>
<th>0dB</th>
<th>5dB</th>
<th>5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>car</td>
<td>SS</td>
<td>-0.7563</td>
<td>1.0078</td>
<td>3.1533</td>
</tr>
<tr>
<td>car</td>
<td>PSS</td>
<td>-1.6769</td>
<td>-0.2694</td>
<td>0.5995</td>
</tr>
<tr>
<td>car</td>
<td>MMSE</td>
<td>-1.2915</td>
<td>0.1892</td>
<td>1.0737</td>
</tr>
<tr>
<td>car</td>
<td>RPCA</td>
<td>-0.2614</td>
<td>1.6150</td>
<td>3.1584</td>
</tr>
<tr>
<td>train</td>
<td>SS</td>
<td>-1.8234</td>
<td>0.4366</td>
<td>3.4502</td>
</tr>
<tr>
<td>train</td>
<td>PSS</td>
<td>-2.3362</td>
<td>-0.8073</td>
<td>1.2418</td>
</tr>
<tr>
<td>train</td>
<td>MMSE</td>
<td>-2.6426</td>
<td>-0.9761</td>
<td>1.4981</td>
</tr>
<tr>
<td>train</td>
<td>RPCA</td>
<td>-0.0133</td>
<td>2.5550</td>
<td>3.386</td>
</tr>
<tr>
<td>street</td>
<td>SS</td>
<td>-0.2393</td>
<td>0.4877</td>
<td>3.3694</td>
</tr>
<tr>
<td>street</td>
<td>PSS</td>
<td>-0.9663</td>
<td>-1.2555</td>
<td>1.5942</td>
</tr>
<tr>
<td>street</td>
<td>MMSE</td>
<td>-1.1467</td>
<td>-1.0628</td>
<td>1.4591</td>
</tr>
<tr>
<td>street</td>
<td>RPCA</td>
<td>-0.3107</td>
<td>1.6997</td>
<td>2.9382</td>
</tr>
<tr>
<td>station</td>
<td>SS</td>
<td>-0.7395</td>
<td>0.9642</td>
<td>3.4255</td>
</tr>
<tr>
<td>station</td>
<td>PSS</td>
<td>-1.7591</td>
<td>-0.3909</td>
<td>0.9407</td>
</tr>
<tr>
<td>station</td>
<td>MMSE</td>
<td>-1.7307</td>
<td>0.1381</td>
<td>1.4394</td>
</tr>
<tr>
<td>station</td>
<td>RPCA</td>
<td>-0.9341</td>
<td>1.8642</td>
<td>3.0396</td>
</tr>
</tbody>
</table>

References