IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 2, April, 2022, pp. 633-644
Journal homepage: http://iieta.org/journals/ts

Speech Enhancement for Robust Speech Recognition Using Weighted Low Rank and Sparse Decomposition Models under Low SNR Conditions

Venkata Sridhar K*, Kishore Kumar T

Department of Electronics and Communication Engineering, National Institute of Technology Warangal, Telangana 506004, India

Corresponding Author Email: sridhar@nitw.ac.in

https://doi.org/10.18280/ts.390226

Received: 14 March 2022
Accepted: 15 April 2022

Keywords:
low rank - sparse matrix decomposition, speech enhancement, nuclear norm, RPCA, SSGODEC WNNM, WSNM, automatic speech recognition, word error rate

ABSTRACT

Noise estimation is a crucial stage in speech enhancement (SE), and it commonly necessitates the use of prior models for speech, noise, or both. Prior models, on the other hand, can be ineffective in dealing with unseen nonstationary noise, especially at low signal to noise (SNR) levels. This paper proposes to assess the efficacy of an unsupervised SE approach based on weighted low rank and sparse matrix factorization to estimate noise and speech when neither is available beforehand by decomposing the input noisy spectrum into a low-rank noise component and a sparse speech component. Due to the approximation of the actual rank of noise, these techniques are constrained, and they do not directly exploit the low-rank property in optimization. Nuclear norm minimization (NNM) is the most well-known approach, as it can precisely recover the matrix's rank under certain restricted and theoretical guarantee conditions. NNM, on the other hand, is unable to reliably estimate the matrix rank in many situations. Significant advancements in computer vision and machine learning applications have demonstrated that a weighted nuclear norm minimization (WNNM), overcomes NNM shortcomings, and achieves a superior matrix rank approximation than NNM. Consequently, in this study, we present alternate SE algorithms that make use of weighted low rank and sparsity constraints to separate speech and noise spectrograms. Following that, they were trained and evaluated on a standard Automatic Speech Recognition (ASR) engine to lower the Word Error Rate (WER). Extensive investigations on the impact of real-world noise on speech signals show that the proposed model outperforms the existing state of art models in terms of objective measures like SDR, PESQ, SIG, BAK, OVL, and STOI values in varied noise circumstances under low SNR environments.

1. INTRODUCTION

The presence of background noise degrades the performance significantly in many speech signal processing applications including mobile communications, man-machine interfaces or hearing aids, low-quality audio devices, ASR systems, etc., Hence, they suffer from reduced speech quality and intelligibility which make their communication troublesome and therefore limit their use. However, for reliable speech communication, ASR in real-world environments must be robust to significant levels of noise. Processing the noisy speech data with each algorithm and retraining the ASR engine for each is impractical in general. This is especially true when a wide range of acoustic situations must be taken into account. One strategy to achieve improved accuracy and robustness is through speech enhancement. In such situations, Speech Enhancement (SE) aims to improve the perceptual quality of speech by separating speech from noisy observations to help to increase recognition accuracy. There is no prior information about the noise and spatial information available in the monaural case. It becomes a major challenge to devise an effective SE strategy, especially under low SNR conditions and with various non-stationary noises frequently experienced in real-life situations.

Over the past several decades, many SE algorithms for monaural recordings have been proposed in the literature to enhance noisy speech signals. The traditional algorithms include spectral subtraction [1], minimum mean square error (MMSE) estimation [2], log minimum mean square error (logMMSE) [3], and Wiener filtering (WF) [4]. These approaches do not require any prior knowledge of speech or noise signals nor any kind of training, therefore they can be used in a variety of contexts. However, because these algorithms assume the noise is stationary, they are ineffective at dealing with nonstationary noise, especially at low SNR. A contemporary and alternate technique is signal subspace decomposition methods [5-7]. Assuming a low-rank linear model for speech and uncorrelated additive noise interference, the decomposition is carried out. Enhancement of speech is accomplished in the temporal domain by masking the noise subspace and then estimating the clean speech signal from the remaining signal-plus-noise subspace. Single-channel SE systems traditionally make use of the Voice activity detection (VAD) stage to estimate and update the noise statistics during noise-only segments. A well-designed VAD will improve the performance of the SE method in noisy environments in terms of accuracy and speed, otherwise, it would degrade the system performance. In low SNR conditions, the current VAD approaches are imperfect. Furthermore, even if VAD is adequately built, alterations in the noise spectrum that occur

during active speech segments are unable to affect the noise estimate in a timely manner. This would lead to an underestimation during long-spoken phrases where there are few noise-only portions [8]. Several algorithms are designed to improve the speech quality by estimating and reducing background noise power spectral density (PSD) for stationary or slow varying noise signals with SNR above 0dB. Although these methods can enhance the speech quality without any prior information about noise type, limited progress is made to improve the intelligibility under unseen non-stationary noise conditions that cannot guarantee a sufficient noise estimation for all scenarios.

To overcome this limitation, many source separation and dictionary learning [9] based noise reduction algorithms have been reported. Such methods include principal component analysis (PCA) [10], Independent component analysis [11], K-SVD [12] and Non-negative matrix factorization (NMF) [13]. Dictionary learning methods play a successful role in machine learning for SE, where the best data vectors (atoms) are modeled as sparse linear combinations of basis factors (dictionary) [14]. While PCA is highly sensitive to noise, corruptions in the data, can estimate the low-rank component arbitrarily far from the true model. Recently, methods like supervised learning using Gaussian mixture model or deep neural network offered better performance with mask estimate. For supervised separation, all these methods always need either a particular feature or huge initial training. However, the discrepancy between diverse real-world noises and training noises will result in performance degradation. To solve this problem, another very elegant remedy theory, called robust principal component analysis (RPCA) [15, 16] was proposed. This is an unsupervised method solved via Principal Component Pursuit (PCP) [17] that decomposes a matrix into low-rank and sparse structures using convex optimization. Simultaneously convex relaxation of the rank minimization model, the nuclear norm minimization (NNM) problem is attracting significant research interest in the recent years.

One can improve perceived audio quality and/or intelligibility with low signal distortion by utilizing the most successful algorithm. Instrumental measures have been developed since human listener testing is time-consuming and expensive. These metrics are designed to estimate how effectively new algorithms will perform by modeling human responses [18-20].

The goal of this research was to develop methods for SE in low-SNR contexts, specifically in cases where a third-party ASR engine is provided either as embedded software or as a cloud-based solution, which could not be adjusted. Many alternative parameters are evaluated and optimized during algorithm development to lower the WER.

1.1 Low-rank and sparse matrix decomposition

From the basic principle of RPCA, the noisy speech spectrum is decomposed into low-rank and sparse matrices using Principal Component Pursuit (PCP) model. The sparse and low-rank components can be approximated and retrieved with a high probability by utilizing efficient estimation procedures. The low-rank matrix approximation (LRMA) method seeks to retrieve the underlying low-rank matrix by minimizing the rank of its relaxations from its corrupted observations of speech. Unfortunately, rank minimization is an NP-hard problem with no known efficient solution. The nuclear norm, which contributes to NNM-based approaches [21], is the best choice for substituting the rank function with its tightest convex relaxation.

The classical Low-rank matrix Factorization (LRMF) method also known as the SVD technique, is capable of achieving the optimal rank-r approximation of input data matrix M by using a truncation operator on its singular value matrix in terms of F-norm fidelity loss. To suppress outliers mixed in data, a robust LRMA method called robust principal component analysis (RPCA) framework, based on nuclear norm minimization (NNM) is introduced. The NNM could be solved by the singular value thresholding algorithm [22] using the alternating direction method of multipliers (ADMM) [23] framework, which also belongs to the augmented Lagrange multipliers (ALM) framework. In the time-frequency (T–F) domain, noise signals present in different time-frames have similar spectral structures and patterns are usually correlated with one other and that can be captured with a few basis vectors. Therefore, the noise spectrogram is supposed to lie in a low-rank subspace. On the other hand, as the spectral energy centralizes in a few T-F units, speech signals can be assumed to be relatively sparse in T–F domain [24]. The RPCA method is a non-parametric method and do not require any specific assumptions about the distribution of the spectral components of either speech or noise. Because both speech and noise spectra can be recovered simultaneously, therefore the process of VAD is unnecessary and irrelevant in this framework. This method is superior to many traditional SE algorithms that depend on the performance of noise estimation algorithms [25, 26]. The RPCA algorithm has the advantages of few tuning parameters and fast processing speed. Moreover, it can perform well in strong noise conditions. This favor to denoise speech through mask estimate on spectrogram via sparse and low-rank decomposition. Sharing similar principles several modifications have been investigated, to improve further the performance of low rank and sparse models like the SS-GoDec [27] algorithm for the SE.

The most noticeable work is nuclear norm minimization (NNM), which can recover the rank of the matrix exactly under some restricted and theoretical guarantee conditions. However, Nuclear Norm Minimization (NNM) based RPCA and SS-GODEC methods may result in undesirable outcomes when prior knowledge of the signal source is not utilized. The standard NNM regularizes each singular value equally, resulting in the simple calculation the of convex norm. This restricts its ability and flexibility in dealing with a wide range of practical challenges in which the singular values have clear physical meanings and should be handled accordingly. Also, these algorithms are limited due to the approximation of the original rank of noise through NNM, and do not explicitly use the low-rank property in optimization. Therefore, for many real-world applications, NNM is not able to approximate the matrix rank accurately, since it often tends to over-shrink the rank components. To rectify the weakness of NNM, recent advances have shown that weighted nuclear norm minimization (WNNM) had shown to achieve a better matrix rank approximation than NNM, which heuristically set the weight as inverse to the singular values. It is proved that the recently proposed WNNM can replace the traditional nuclear norm, as an improved approximation to the rank of a matrix in computer vision applications [28]. As RPCA and SS-GODEC algorithms explicitly account for deviations of the speech and noise time-frequency matrices from the idealistic sparse and low-rank model, we propose an alternate SE algorithm for speech and noise spectrogram separation by imposing

weighted low rank and sparsity constraints. With the help of the low rankness of WNNM, the efficacy of enhancement by using singular value decomposition, the ADMM, and the accelerated proximal gradient line search method is improved. Therefore WNNM-based RPCA enhancement model is proposed here which takes the advantage of the high correlation of the speech signals, showing excellence to the NNM-based methods. Further study led to the invention of a new RPCA model, the weighted Schatten p-norm minimization model, to effectively perform low-rank regularization (WSNM). It is demonstrated in [29] that WSNM suppresses noise more effectively than state-of-the-art approaches and is better at modeling dynamic and complex situations. WSNM is a generalized version of WNNM whose performance in image denoising was analyzed.

Motivated by this, the work evaluates and presents the results of NNM based RPCA, SS-Go Dec, WNNM based RPCA, and WSNM for enhancement of speech signals. The best measure for evaluating the performance of SE algorithms for ASR is the WER achieved over a specific set of data. In the first experiment, the performance of SE algorithms on a diverse range of acoustic conditions is evaluated using BSS-eval metrics. These evaluations are conducted on the standard set of test speech signals that include: NOIZEUS [30], TIMIT [31], and LIBRI [32] databases. Substantial experiments conducted exploit the influence of these algorithms for a wide range of nonstationary real-world noise conditions. The induction of masks in Speech Enhancement has shown a lot of promise in improving speech intelligibility. Therefore, binary T-F masks and log-sigmoid soft masks are also considered for the enhancement algorithm and experimented with it. Further, the results are analyzed to investigate the suitability of the low-rank and sparse model for speech and noise signals. Finally, the performances of the baseline SE methods like KSVD, NMF are compared with RPCA, SS-Go Dec, WSNM, and WNNM models.

The purpose of the second experiment is to determine the effectiveness of different realizations proposed for SE under low SNR with least WER using the standard kaldi ASR system [33]. For mobile communicating devices, in particular, the variations in performance due to a wide range of acoustic conditions are taken into account. The results illustrate that the performance of our proposed approach is better than those of existing state of art methods.

With the proposed model under low SNR conditions, promising results were obtained in our experiments in terms of better objective measures such as: SDR, PESQ, SIG, BAK, OVL, and STOI values when compared with baseline methods. The important thing to note from the findings is that for all types of noises and all masks, the proposed approach achieves an output SDR that is significantly higher than the input SDR. Specifically at -10dB input SNR level, using the WSNM model, an improvement of 8.14 dB and 6.17 dB in output SNR is observed with Traffic & Car and Wind noise, respectively. For the same settings PESQ scores of 2.51 and 2.27 respectively were achieved. For input noise specifically between -10 to 0 dB, it is observed that the proposed WSNM algorithm with a binary T-F mask increased speech intelligibility. In all noise situations, the trained ASR with libri speech corpus performed effectively in low SNR levels (< 0 dB) and significantly decreased WER when compared to baseline techniques. Overall, the results demonstrate that the performance of our proposed approach is better than those of existing state of art methods.

The paper is organized as follows: Section 2 provides the overview of the SE Methods using RPCA, SS-Go Dec. It also contains algorithmic frameworks for implementations of these matrix decompositions. Section 3 describes the proposed WNNM, and WSNM based SE algorithm step by step and explains the STFT process and the time-frequency masking process. All information about the experimental setup and the methodology for this work, results, and analysis of these results are presented in sections 4 and 5. Section 6 presents the discussions and conclusions.

2. OVERVIEW OF RPCA FOR SPEECH ENHANCEMENT FRAMEWORK

RPCA breaks down noisy speech into low rank and sparse components in T-F domain based on convex optimization. The benefits of adopting RPCA are expressed by the fact that non-stationary noise is frequently less spectrally diverse than foreground speech in SE. Noise signals are assumed to be low-rank components because their spectrograms in time frames are correlated, but speech signals are considered to be sparse components due to their sparseness in the frequency domain. Based on alternating projection algorithm, speech and noise magnitude spectrograms are subjected to sparsity and rank constraints in order to enhance noisy speech [34]. In this work, the RPCA based methods will be evaluated regarding their ability to enhance speech signals. The RPCA algorithm first performs an STFT via the overlap-add method on the noisy speech input signal X_{NS}. Then RPCA is solved via PCP, decomposes magnitude matrix $|X_{NS}| \in Rn1\times n2$, i.e. the spectrogram, into low-rank L and sparse S components, where $L \in Rn1\times n2$ and $S \in R\ n1\times n2$, While the phase information is stored for the reconstruction later on. X_{mixed} contains the spectrograms of a pair of original speech and noise contributions.

Next, as in the studies [7, 8], L and S are used to create a mask: $M \in Rn1\times n2$ with $0 \leq M(m, n) \leq 1\ \forall\ m,n$.

The final noise and speech estimates $|\hat{X}_{noise}| \in Rn1\times n2$ and $|\hat{X}speech| \in Rn1\times n2$ are calculated with the help of time-frequency (T-F) mask. This T-F masking step was taken into consideration as a way to possibly further improve the enhancement results. The time-domain estimates $\hat{x}_{speech}[k]$ and $\hat{x}_{noise}[k]$ are then calculated via overlap & add inverse STFT. The separation that is performed by this RPCA driven algorithm is based upon recognizing the clean speech by its sparsity and identifying the noise by its low-rank character. That is, the spectral structure of the speech component varies quickly with time while the noise component is usually either fixed or varies slowly. Such property signifies that clean speech is sparse while noise part appears to be low rank [35]. Therefore extracting the sparse component in the noisy speech matrix tends to enhance the noisy speech by reducing the noise. While the accuracy of the low-rank assumption for the noise will depend to great extent on the kind of noise that is present in the signal. A stationary and repetitive noise, like the noise of a machine, for example, would much rather be assumed to exhibit a low-rank character than a dynamic and varying crowd noise.

However, it is observed from several experiments that the sparsity assumption seems to be appropriate for the speech signal. Assessing the suitability of the low-rank and sparse model and its influence on the enhancement results will be an important aspect of the evaluation in this work and will be

done in Section 4.2. The results that will be presented there do indeed show that the low-rank assumption for a certain type of noise signal is problematic. Many different T-F representations are depending on the combination of parameters that are chosen for the calculation of the STFT. The STFT length can be adjusted and the STFT window type can be chosen. Further on, the hop size can be selected, i.e. the number of samples by which the window is shifted between two consecutive FFTs.

However, for all reasonable choices of the parameters, the resulting spectrogram matrix together with the associated phase matrix contains all the information about the transformed signal and the signal can exactly be reconstructed via ISTFT. Despite this, different ways of calculating the STFT have different effects on the success of the SE algorithm. The effects of different STFT settings that were found in the results of the experiments for this paper are described in detail in Section 4.1

2.1 SE method using NNM based RPCA

RPCA aims to find sparse version and a low-rank version from a noisy speech data matrix [35, 36]. It is robust to outliers and therefore finds extensive applications in image enhancement, modeling of images, separation of background in images, videos, and different machine learning purposes. There are several algorithms to solve RPCA problem such as: PCP, Inexact ALM (IALM), ADMM etc. RPCA solved via PCP decomposes a data matrix M, represented in Eq. (1) as follows:

$$M=L_0+S_0 \tag{1}$$

where, L_0 is the low-rank matrix and S0 is the sparse matrix.

The RPCA is calculated using Eq. (2) as follows:

$$\arg\min (||L||_* + \lambda ||S||_1) \text{ under the constraint } M=L+S \tag{2}$$

where, λ is a positive constant that regulates the relative weight between the rank minimization and l_0-norm.

The approximated values of $\hat{L}$, $\hat{S}$ are calculated using Eq. (3) as follows:

$$\hat{L}, \hat{S} = arg\ min_{L,S} \left|\left|L\right|\right|_* + \lambda\left|\left|S\right|\right|_1 \text{ s.t M-L-S = 0} \tag{3}$$

By using the Lagrange method, a Lagrange multiplier Y is associated to produce an unconstrained function. The optimum values of L and S are found in an iteration using the Y value from the last iteration. Thus, in this way, the values of L, S, and Y are updated to reach the global optimum.

2.2 SE method using SS GODEC

From the preliminary experiments, it is observed that the RPCA model is not effective and robust to extract the formant structure of original speech. Thus, RPCA method is modified to use the GO-Dec algorithm by representing spectrogram of the real-world noisy speech M as the superposition of L, S, and E, that is, M = L+ S + E, where L and S are the low-rank and Sparse components and E is a noise term that perturbs the ideal low-rank and sparse character. Following this, the optimization objective function is formulated in Eq. (4) as follows:

$$\arg min_{L,S} \|M - L - S\|^2 \; s.t\; rank(L) \le r \;\; and\; card(S) \le m \tag{4}$$

And yields low-rank and sparse estimates $\hat{L}$ and $\hat{S}$ [19]. So, L and S have to be chosen such that they meet the predefined conditions on their rank and cardinality of their support set while the noise power $\|E\|^2 = \|M - L - S\|_F^2$ is minimized.

As the cardinality S is hard to estimate, by using soft threshold λ for matrix decomposition, the optimization problem is formulated in Eq. (5) as follows:

$$arg\ min \|M - L - S\|^2 + \lambda\|S\|_1 \; s.t\; .rank\ (L) \le r \tag{5}$$

3. RPCA BASED WEIGHTED NUCLEAR NORM MINIMIZATION (WNNM) FOR LOW-RANK ESTIMATION

The goal of NNM decomposition is to recover the underlying low-rank matrix L from its degraded observation matrix M, by minimizing ||L||∗. But the main problem with the above formulation of NNM-RPCA is that the optimization function is non-convex and the problem falls under NP-hard problems, which are computationally expensive. Moreover, the technique assigns equal weights to all the singular values or rank components resulting in biased estimate of low rank and sparse components, restricting its flexibility in practical applications. The singular values of a matrix in the context of speech processing are closely associated with the physical properties of the speech signal. Large singular values account for prominent features of speech such as short-term zero crossing and energy, whereas smaller ones correspond to noise components. Therefore, large singular values must be treated differently from the smaller ones and must be preserved to reproduce high-quality speech. To improve the performance of NNM, in the last few years, numerous applications based on NNM have been proposed, such as video enhancement, background extraction, and subspace clustering. However, the nuclear norm is generally adopted as a convex surrogate for matrix rank. The singular value thresholding (SVT) model for NNM treats different rank components equally, leading to over shrink the rank components, and hence the estimation of the matrix rank is inaccurate. As a result, it is obvious that the traditional NNM model, as well as the accompanying SVT approaches, are insufficiently adaptable to deal with such problems. The methods such as truncated nuclear norm regularization (TNNR) and the partial sum minimization (PSM] among N singular values, keep the largest 'r' (rank of the matrix) singular values unchanged and only minimize the smallest (N-r) ones. TNNR and PSM, on the other hand, are not flexible enough because they make a binary decision on whether or not to regularize a particular singular value or not. While could produce an over-fitting solution due to the noise effects.

Inspired by the singular values that have distinct physical implications proposed the weighted nuclear norm minimization (WNNM) model. WNNM generalizes NNM and improves the flexibility of NNM significantly. To improve the flexibility of nuclear norm, in this work we propose to investigate the weighted nuclear norm and evaluate its minimization strategy. The weighted nuclear norm of a matrix M is defined in Eq. (6) as follows:

$$||\ M\ ||_{w,*} = (\Sigma \| w_i\ \sigma_i(M)\|_1 \tag{6}$$

where, vector w = [w_1, w_2 ..., w_n] and $w_i \geq 0$ is a non-negative weight assigned to σ_i. The rational weights rules for weighting can be specified depending on the prior knowledge and understanding of the problem, which will greatly improve the representation capability of the original data from the corrupted input. From prior knowledge, it is understood that the higher singular values of M are more essential than the smaller ones in natural speech because they indicate the energy of the major components of M. The larger the individual values are, the less they should be shrunk while denoising. As a result, it's a natural assumption that the weight given to σ_i(M), i-th singular value of M, should be inversely proportional to σ_i (M). WNNM is a non-convex problem that is more complex to solve than NNM. So far the WNNM problem has got very little attention in this work. We investigate in depth the WNNM problem using F-norm data fidelity. The solutions are examined under various weight conditions.

We implement the proposed WNNM algorithm to SE as a significant application. SE aims to estimate the hidden clean speech from its noisy observation. As a classical and fundamental problem in low SNR conditions, SE has been extensively explored for many years; however, it remains a prominent research area since enhancement is an ideal testbed for investigating and evaluating the statistical speech modeling techniques. The use of speech Nonlocal self-similarity (NSS) has improved significantly the SE performance in recent years. The NSS prior refers to the fact that for a given local frame in a natural speech, one can find many similar frames to it across the speech signal. The nonlocal similar frame vector is stacked into a matrix, which must be a low-rank matrix with sparse singular values. As a result, enhancement algorithms can be designed using low-rank matrix approximation method. This research provides a two-fold contribution. First, we examine the WNNM optimization problem in-depth and propose solutions for various weight conditions. Second, we demonstrate the potential of the proposed WNNM algorithm for SE in low SNR situations.

3.1 Problem formulation for WNNM model

RPCA attempts to identify a low-rank version and a sparse version from a single matrix and has a wide range of applications. In this section, we propose reformulating Eq. (1) using the weighted nuclear norm, resulting in the WNNM based RPCA (WNNM-RPCA) model represented in Eq. (7) as follows:

$$\begin{aligned} &\arg\min \left(||L||_{w,*} + \lambda ||S||_1 \right) \\ &under\ the\ constraint\ M = L + S \end{aligned} \tag{7}$$

The ADMM is used to solve the WNNM-RPCA problem, just like it is in NNM-RPCA. By using the ALM method, a Lagrange multiplier Y is associated to produce an unconstrained function represented in Eq. (8) as follows:

$$\begin{aligned} arg\ min_{L,S} ||L||_{w,*} + \lambda ||S||_1 + < Y, M - L - S \\ > + \frac{\mu}{2} ||M - L - S||_F^2 \end{aligned} \tag{8}$$

where, μ=1/2k.

The optimum values of L and S are found in an iteration using the Y value from the last iteration. Then again, the value of Y is updated in the current iteration with the new optimum L and S values.

Lk, Yk, and Sk are local variables and represent the local optimum in the kth iteration is represented in Eq. (9) and Eq. (10) as follows:

$$\begin{aligned} S_{k+1} &= argmin\, f(S, L_k, Y_k) \\ &= argmin_S\, \lambda || S ||_1 \\ &\quad + \frac{\mu}{2} || M + (Y_k/\mu) - L_k - S ||_F^2 \end{aligned} \tag{9}$$

similarly

$$\begin{aligned} L_{k+1} = argmin_L\, \lambda || L ||_{w,*} \\ + \frac{\mu}{2} || M + (Y_k/\mu) - L \\ - S_{k+1} ||_F^2 \end{aligned} \tag{10}$$

For the weight wi of each group Mi, large singular values of each frame group mj in M usually offer significant information, and vice versa, inspired by singular values that have clear physical implications. As a result, we usually shrink large singular values less and smaller singular values more. To put it in other words, the weight wi of each group mj in M is set to be inverse to the singular values, and so as in the study [15], the weight is heuristically set as:

$$w_{i,j} = c / (\sigma_{i,j} + \epsilon),$$

where, c and ϵ are the small constants.

Solving the above equation, we obtain Eq. (11) as follows:

$$Y_{k+1} = Y_k + \mu_k (M - L_{k+1} - S_{k+1}) \tag{11}$$

Thus, in this way the values of L, S and Y are updated to reach the global optimum.

Algorithm 1 SE by WNNM-RPCA

Input: Noisy speech data M, weight vector w
1: Initialize $\mu_0>0$, $\lambda>0$, $\rho>1$, $\theta>0$, k=0, L_0=M, Y_0= 0;
2: do
3: $S_{k+1} = argmin_S\, \lambda || S ||_1 + \frac{\mu}{2} || M + (Y_k/\mu) - L_k - S ||_F^2$;
4: for each frame mj in M do
5: Find similar frame group Mj
6: Estimate weight vector w
7: $L_{k+1} = argmin_L\, \lambda || L ||_{w,*} + \frac{\mu}{2} ||M + \left(\frac{Y_k}{\mu}\right) - L_k - S_{k+1} ||_F^2$;
8: $Y_{k+1} = Y_k + \mu_k (M - L_{k+1} - S_{k+1})$;
9: Update $\mu_{k+1} = \rho * \mu_k$;
10: k = k + 1;
11: while $||M - L_{k+1} - S_{k+1}||_F / ||M||_F > \theta$;
12: Output Matrix L = L_{k+1} and S = S_{k+1};

3.2 Problem formulation for WSNM model

WSNM is a generalized variant of Weighted Nuclear Norm Minimization, whose image denoising performance has been studied in Refs. [13, 14]. WSNM Low-rank approximation tends to carry out low rank regularization effectively wherein we employ the loss function expressed in Eq. (12) as follows:

$$arg\ min_{S,L} || S ||_1 + || L ||_{w,S_p}^p\ s.t\ M = L + S \tag{12}$$

Using Augmented Lagrangian function, we get Eq. (13) as follows:

$$L(L,S,Z,\mu) = ||S||_1 + ||L||_{w,S_p}^{P} + < Y, M - L - S > + \frac{\mu}{2}||M - L - S||_2^F \quad (13)$$

where, Y is Lagrangian multiplier, μ is positive scalar. The values of the weighted vectors are defined in Eq. (14) as follows:

$$w_i = C\sqrt{(mxn)}/(\sigma_i(M) + \epsilon) \quad (14)$$

Algorithm 2 SE by WSNM-RPCA

Input: Noisy speech data M, weight vector w, power p
1: Initialize $\mu_0 > 0$, $\rho > 1$, k = 0, $L_0 = M$, $Y_0 = 0$;
2: do
3: $S_{k+1} = argmin_S \lambda||S||_1 + \frac{\mu}{2}||M + \left(\frac{Y_k}{\mu_k}\right) - L_k S_k||_F^2$;
4: $L_{k+1} = argmin_L ||L||_{w,Sp}^{p} + \frac{\mu}{2}||M + \left(\frac{Y_k}{\mu}\right) - L - S_{k+1}||_F^2$;
5: $Y_{k+1} = Y_k + \mu_k (M - L_{k+1} - S_{k+1})$;
6: Update $\mu_{k+1} = \rho * \mu_k$;
7: k = k + 1;
8: while $||M - L_{k+1} - S_{k+1}||_F / ||M||_F$ not converged;
9: Output Matrix $L = L_{k+1}$ and $S = S_{k+1}$;

4. EXPERIMENTAL SETUP AND METHODOLOGY

This section provides an experimental setup and methods for evaluating the suggested noise reduction methods' performance and suitability. The results of these experiments are of high informative value for assessing the possibilities and limitations of the enhancement method. They also allow estimating the influence of the parameters that were considered in Sections 2 and 3. Because of many uncertainties and complicated relations, the theoretical discussion in the previous section did not suffice to make a reliable prediction of the performance of the speech recovery procedure. On the contrary, the results of the experiments show how well the algorithms have already performed in tests, and the assumption is justified that they will perform similarly in identical situations. Therefore, the following section contains very valuable information about the potential of the SE method in practical use. Even more so as the number of test signals that were used is rather high.

The standard Noizeus corpus [30] was used in studies. The speech signals available as wav-files with a sampling rate of 8 kHz. A total of 20 clean sentences were chosen for this study The noisy stimuli were created by adding clean phrases with five different signal-to-noise ratio levels, including -10, -5, 0, 5 and 10 dB. The noise signals obtained from a noise collection available as waveforms with a sampling rate of 8 kHz as well. Five noise recordings: The cheering of a crowd of people, a bubbling stream of water, wind, machine, and car driving in traffic were selected for the evaluation. AWGN is simulated and added to the clean speech. This resulted in an overall number of 5 • 20 • 6 = 600 mixed test signals, which are all about 3 seconds long.

Low-rank, sparse and noise matrix decomposition algorithms needs to be given the parameter r and λ. r determines the rank of the low-rank component while λ is used to trade off the desire to minimize the cardinality of the sparse component against the desire to minimize the energy of the noise component. It is important to tune the two parameter. If λ is chosen too small, then parts of the noise will leak into the speech estimate because the urge to minimize cardinality of the sparse component is not high enough to eliminate all relevant noise contributions from the sparse component. If λ is chosen too big on the other hand, the urge to minimize the cardinality of the sparse component is so dominant that parts of the speech will be eliminated from speech estimate which is counterproductive of course. It should be pointed out that apart from the parameters that were changed in order to evaluate their influence on the performance of the SE method all settings were left as they were. The best value in this investigated is an average output SDR of 2.89 dB which was achieved by setting r = 1 and λ = 1. Therefore, this will be the setting that will be used in the following comparison of the performances in the speech denoising methods.

4.1 Influence of binary and log-sigmoid time-frequency masking

In order to illustrate the time-frequency masking step and the influence of different masks, Figure 1 contains plots of all matrices that are relevant for example masking step. The spectrogram of the noisy speech input signal is shown in Figure 1a. Figures 1b and 1c show the low-rank and sparse components that decomposed the input spectrogram by WNNM-RPCA algorithm. Figure 1d depicts the binary mask derived from the low-rank and sparse components, whereas Figure 1e display the final speech estimate after applying the binary mask. Figure 1f shows the log-sigmoid mask calculated from the low-rank and sparse components and Figure 1g is the final voice estimate after applying the log-sigmoid mask. Figure 1h shows the final speech estimate spectrogram when no mask is applied. The SE algorithm with RPCA matrix decomposition was applied to all 600 test signals. With fixed STFT settings of M = 1024, hop-size = 256 using Hanning windowing, the enhancement methods were tested with binary masks, log- sigmoid masks and without masks in order to experimentally evaluate and compare the influences of the masks.

For all five noise types, all five speech to noise energy ratios in the input signal and all three different masks (no mask, binary mask and log-sigmoid mask), the mean speech to distortion energy ratio is obtained by averaging the resulting SDR values of all 20 speakers. The most obvious and most important thing that can be learned from these results is that for all noises and all that for all noises and all masks, the approach with the settings as specified above achieves a speech to distortion energy ratio (SDR) in the speech estimate that is considerably higher than the speech to noise energy ratio in the input signal.

The most obvious and most important thing that can be learned from these results is that for all noises and all masks, the approach with the settings as specified above achieves a speech to distortion energy ratio (SDR) in the speech estimate that is considerably higher than the speech to noise energy ratio in the input signal. This is true for all entries with input SNR levels of -5dB, 0dB and 5dB and most of the entries with input SNR of -10dB. Only for some entries at the already high input SNR of 10dB, the enhancement method fail to produce a further increase in the speech quality and decreased it instead. From the results, it is observed that all three masks achieve very similar results for low values of the input SNR and that the output SDR values become slightly more spread out for higher values of the input SNR. For high SNR values, the SE algorithm without any mask does perform best, log-sigmoid

masking is second best and binary masking is last with a performance that is about 1dB worse than that without masking. This suggests that the WNNM-RPCA decomposition already achieves a good separation of speech and noise, which cannot be further enhanced with the masks used here. Instead, the masks cause undesired alterations that deteriorate the results.

For low values of the SNR on the other hand, the results are closer together with log-sigmoid masking performing best. So, in very noisy conditions, masking can help improve the outcome of the speech recovery a little bit. Another aspect that results reveal is that log-sigmoid masking does constantly perform about 0.4dB better than binary masking. This is not only true on average but can also be verified by comparing corresponding individual entries. It can be realized that the behaviour of the three different mask types for the individual noise types does not deviate significantly from the average of overall noise signals. This means that none of the tested noise signals has a mask type that is particularly suitable and performs significantly better than all other mask types.

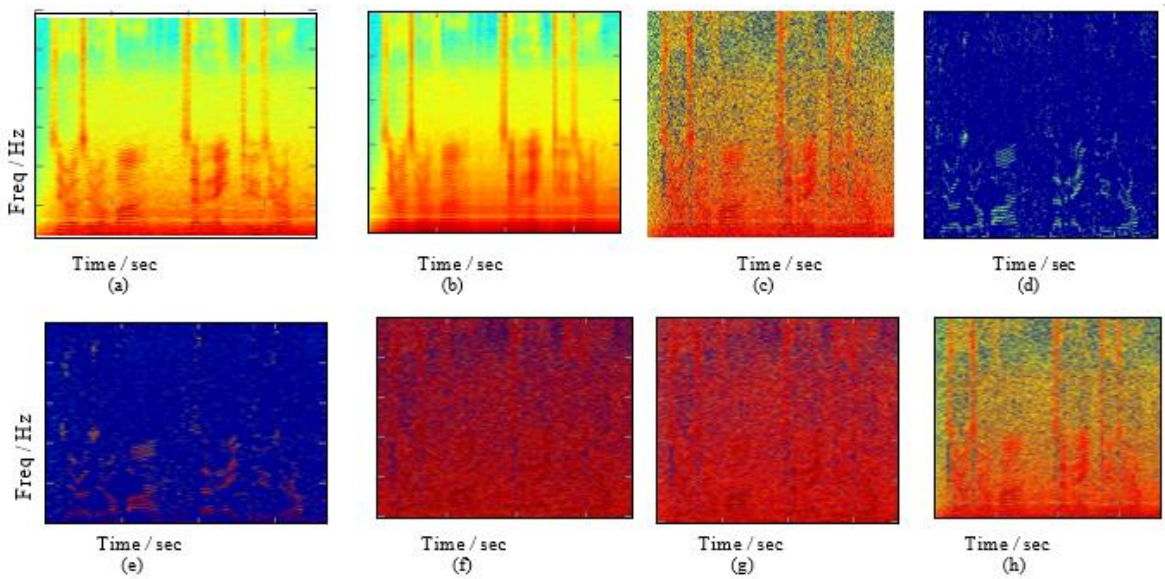


Figure 1. Plots of relevant matrices for the time-frequency masking step using WNNM-RPCA Based SE algorithm: a) Spectrogram of noisy speech signal b) Low-rank component. c) Sparse component d) Binary mask e) Speech estimate after binary masking f) Log-sigmoid mask g) Speech estimate after Log-sigmoid mask h) Speech estimate without masking

Figure 2. Performance comparison of the proposed SE algorithms with baseline methods in terms of SDR values using standard NOIZEUS data base

4.2 Evaluation of the weighted low rank and sparse decomposition models for SE

This section contains the evaluation results that were achieved with the WNNM-RPCA (WRPCA) and WSNM-RPCA (WSRPCA) based enhancement procedure for different settings of the parameters which were discussed in Sections 3.1 and 3.2. The 600 test signals were decomposed with the SE algorithm that uses NNM-RPCA at its core. The results indicate how well the WSRPCA based SE algorithm will perform for different noise types. The suggested SE algorithms are evaluated and validated against the baseline state-of-the-art SE algorithms using objective evaluation metrics such as: SDR, PESQ, STOI, SIG, and BAK. The results of experiments revealed that WSRPCA outperforms state-of-the-art enhancement algorithms not only in terms of PESQ and STOI index but also in local structure preservation, leading to listening more pleasant.

Figure 2 shows the performance comparison of the proposed and baseline SE algorithms in terms of SDR. At -10dB, using the suggested WSRPCA approach, an improvement of 8.14 dB and 6.17 dB in SDR was observed with Traffic & Car and Wind noise, respectively. The weighted low rank and sparse models have shown improvements in all SNR levels and noise environments. The proposed methods are also examined with AWGN as a stationary noise case.

The performance study of the proposed algorithms versus KSVD, NMF, RPCA, and SS-GODEC in terms of PESQ [37] for all SNR levels is depicted in Figure 3. PESQ was improved the most in noisy unprocessed speech at -10 dB traffic and car noise (ΔPESQ = 0.49) and the least with 10 dB AWGN (ΔPESQ = 0.27). When compared to the baseline techniques, the suggested speech enhancement algorithms showed a considerable improvement in PESQ at all SNR levels and noise situations. At -10 dB noise levels, the greatest PESQ scores were obtained in traffic and car noise, wind noise, and AWGN, with PESQ = 2.51, 2.27, and 2.43, respectively.

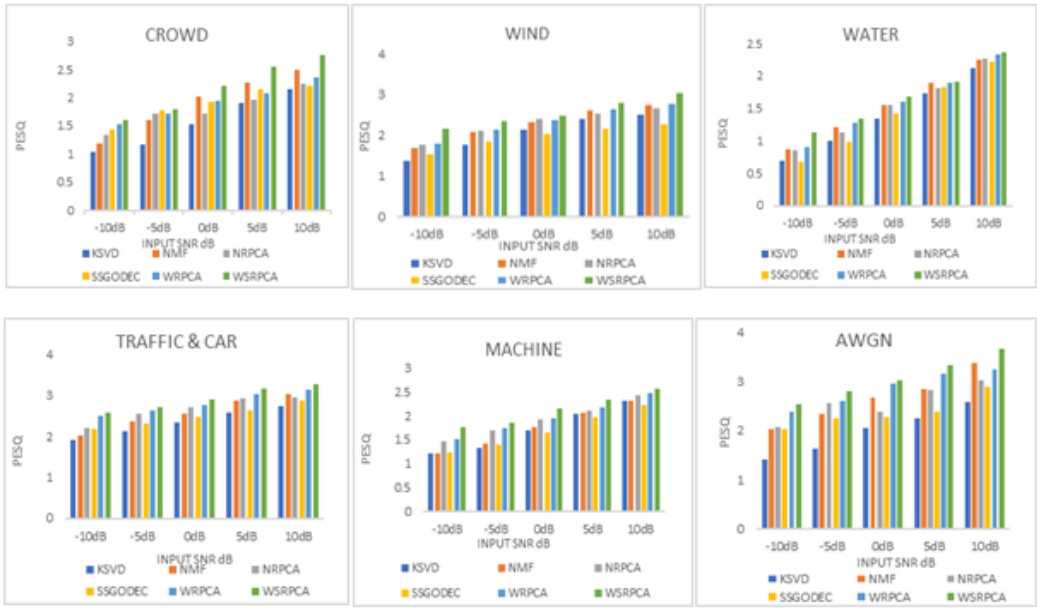


Figure 3. Performance comparison of the proposed SE algorithms with existing methods in terms of PESQ values using standard NOIZEUS data base

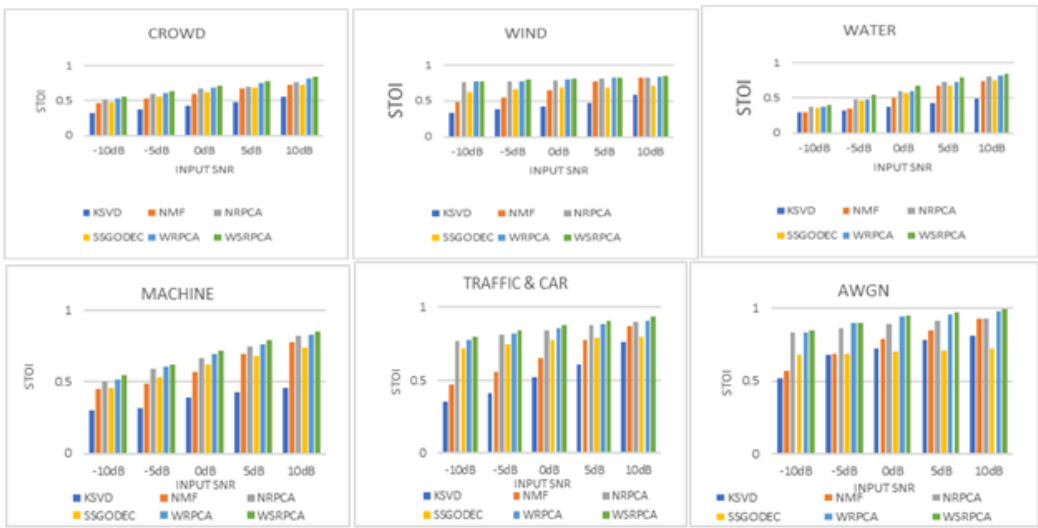
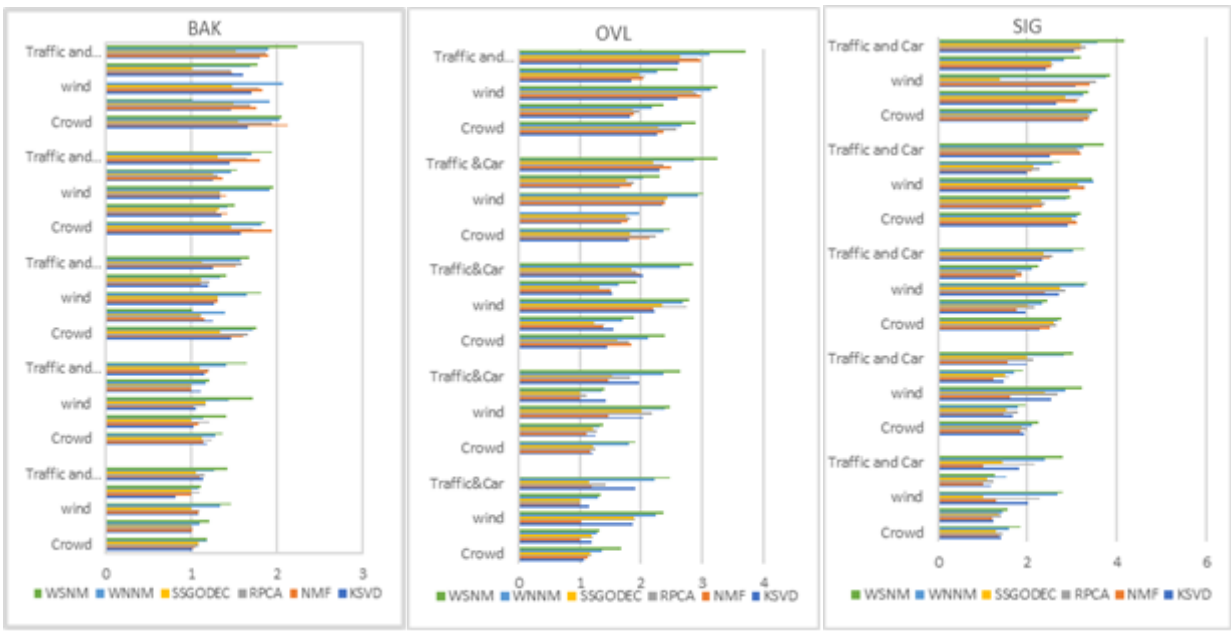


Figure 4. Performance comparison of the proposed SE algorithms with existing methods in terms of STOI using standard NOIZEUS data base



(a) (b) (c)

Figure 5. Performance comparison of the proposed SE algorithms with existing methods in terms of Objective metrics: a) BAK b) OVL c) SIG

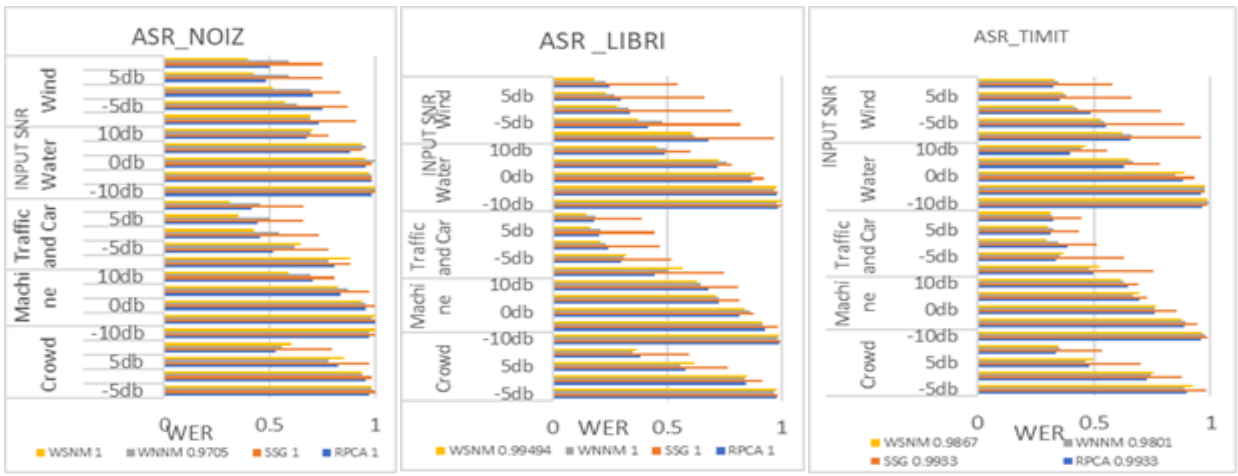


Figure 6. Performance comparison of proposed speech enhancement algorithm in terms of Word Error Rate (WER) over noisy and baseline algorithms

According to the results of the previous investigations on proposed algorithms, using a binary T-F mask improved speech intelligibility in strong noisy conditions (-10 to 0 dB). Figure 4 demonstrates the improved speech intelligibility with binary mask using STOI [38] measure. For SNR = 10 dB, all noise sources resulted in the highest intelligibility scores (STOI > 0.86).

To determine the speech distortion and back ground residual noise introduced by the recommended algorithms, measures like SIG, BAK, and OVL must be taken into account. For the proposed algorithm's speech distortion (SIG), residual noise

(BAK) and Overall quality with similar are measured and are shown in Figures 5 (a-c). The offered strategies provides low residual noise and consistently produces high BAK and OVL values at all SNR levels and noise conditions. In all noise situations, the algorithm performed effectively in low SNR levels (-10 dB) and significantly decreased residual noise when compared to baseline techniques. The proposed approach produces high SIG values at all SNR levels and noise situations, demonstrating its usefulness in terms of speech content preservation.

When compared to baseline techniques, in low SNR situations (< 0dB) the proposed approach produced the greatest SIG values, demonstrating its usefulness in terms of speech content preservation. The approach outperformed in Traffic & car noise at all SNR levels by introducing less speech distortion and also very little residual noise in all noise settings.

5. EXPERIMENT RESULTS ON ASR

According to WER at various input SNR's, the human ability to recognize speech contents remains resilient. However, in the presence of strong background noise, the accuracy of single channel ASR systems decreases significantly. In fact, most single-channel speech enhancement (SE) approaches (denoising) have only provided marginal performance improvements over state-of-the-art ASR back ends trained on multicondition training data. One of the best

ways to improve robustness to a speech recognition system is to include a noise reduction (SE) stage.

Testing the performance of each SE algorithm throughout the complete spectrum of acoustic circumstances takes a long time. As a result, it is preferable to estimate WER scores using more easily computed metrics during the development of the SE algorithm, where the clean speech reference is available. Predicting the performance of the SE algorithm is beneficial to correlate the improvements in WER with improvements in bss_eval metrics. The denoised speech signals from the first experiment are used in the second experiment to test and evaluate WER employing the kaldi ASR repository. Among the speech processing schemes experimented, particularly the proposed speech enhancement algorithms performed well in terms of the WER over noisy speech, depicted in Figure 6.

6. DISCUSSIONS AND CONCLUSIONS

This paper proposes two convex optimization-based speech enhancement approaches that don't require any prior knowledge of speech or noise. Using a low rank sparse matrix decomposition model, the approach decomposes the input noisy speech magnitude spectra into low rank noise and sparse speech components. We feel that the provided algorithms will be a new feasible direction for the SE problem under low SNR conditions due to their superior features. The suggested methods are non-parametric strategies that do not require any assumptions about the spectral component distribution in speech or noise. In the T-F domain, it only requires low-rank noise and sparse speech. The VAD approach is irrelevant and unnecessary in this SE framework since speech and noise components can be obtained simultaneously [39, 40].

The contribution of this research is to provide an unsupervised speech denoising strategy under diverse, strong, and unseen real-world nonstationary noisy settings that uses low rank and sparse decomposition models with a different objective function than the conventional RPCA approaches. For each noisy input, all the regularization parameters are automatically modified and updated. Although the existing methods such as KSVD and NMF methods can eliminate most interferers, under low SNR conditions (< 0dB) part of the recovered speech formant structures are lost during the matrix decomposition process, resulting in speech distortion. To alleviate speech distortion, we intend to build a novel low-rank and sparse matrix decomposition model by placing appropriate constraints on the sparse part. The present study assessed several objective measures widely used for evaluating speech quality.

In the first experiment, we have evaluated the performance metrics of RPCA, SS-GODEC, WNNM and WSNM and compared with KSVD and NMF in a wide range of acoustic conditions. The test conditions included speech signals from the Noizeus data bases and five real world noise at five SNR levels (-10 dB, -5dB, 0 dB, 5 and 10 dB). Acoustic conditions with stationary noise at various SNR levels are included in our experiments that has shown excellent performance. With the proposed model, promising results have been obtained in our experiments in terms of better objective measures like SDR, PESQ, SIG, BAK, OVL, and STOI values when compared with baseline methods such as KSVD, NMF, RPCA and SS-GoDEC.

Second experiment is conducted to investigate and compare the presented SE algorithms with the RPCA models for speech recognition. In this, enhanced speech signals from the first experiment are trained and evaluated WER using kaldi ASR repository. We examined the generalization capability of SE methods using Noizeus, Libri, TIMIT data bases and ASR backends. The ASR results shows that the performance of our proposed approach with libri database produced the lowest WER values.

SS-GoDec algorithm is determined to have a negative impact on WER when compared to other algorithms studied. The solutions are investigated under various weight conditions of nuclear norm, and the proposed WNNM and WSNM algorithms outperforms the NNM problem. Since the algorithms tested here use standard data bases, any change in the spatial properties in the acoustic channel (closed room) may likely degrade the performance.

Some thoughts should be elaborated here in order to conclude the presented evaluation of the influence that different STFT parameters have on the results of the RPCA based SE methods. First of all, it should be emphasized that the presented studies of the effect of changes in the STFT settings are of course not at all exhaustive. Numberless values for each parameter could be tested and for an infinite number of combinations of parameters the performance of the SE algorithm could be evaluated. Absolutely remarkable is the performance of the WSNM-RPCA based unsupervised and untrained approach is successful even when applied for challenging unsteady noises such as the sound of a bubbling stream of water, people cheering of the crowd of people. The improvement can be achieved even under very noisy conditions with low input SNR values. Then, even more, extensive test signal corpora and more elaborate objective speech quality measures could be used. It should also be investigated further which noise types the performance of the SE method tends to be good for and which noise types are too challenging. The good results of the WSNM based RPCA approach for some noise types could motivate the development of real-time realizations of this algorithm, which could be interesting for hands-free mobile communication in cars or hearing aids, for example. At low SNR levels, the proposed algorithm offered minimal speech distortion, and little residual noise was found in speech processed by the proposed algorithm, as shown by high BAK and SIG values, respectively. The spectrogram and time-domain analysis further revealed that the suggested algorithm's output speech had low residual noise. Our study demonstrated that the use of convex optimization methods like: WNNM and WSNM has greatly improved the performance of SE under low SNR conditions.

The proposed SE methods, however, are unable to completely remove background noise since the convex optimization techniques are inaccurate in estimating exact low rank. The problem of developing robust speech enhancement algorithms that can effectively remove background noise yet maintaining good quality and intelligibility in highly nonstationary and adversely noisy situations has yet to be solved. For Superior performance, models to estimate exact low-rank and noise type are to be explored.

REFERENCES

[1] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2): 113-

120. https://doi.org/10.1109/TASSP.1979.1163209
[2] Ephraim, Y., Malah, D. (1984). Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(6): 1109-1121. https://doi.org/10.1109/TASSP.1984.1164453
[3] Ephraim, Y., Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 33(2): 443-445. https://doi.org/10.1109/TASSP.1985.1164550
[4] Liutkus, A., Badeau, R. (2015). Generalized wiener filtering with fractional power spectrograms. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 266-270. https://doi.org/10.1109/ICASSP.2015.7177973
[5] Ephraim, Y., Van Trees, H.L. (1993). A signal subspace approach for speech enhancement. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 355-358. https://doi.org/10.1109/ICASSP.1993.319311
[6] Hu, Y., Loizou, P.C. (2003). A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Transactions on Speech and Audio Processing, 11(4): 334-341. https://doi.org/10.1109/TSA.2003.814458
[7] Hermus, K., Wambacq, P., Van hamme, H. (2006). A review of signal subspace speech enhancement and its application to noise-robust speech recognition. EURASIP Journal on Advances in Signal Processing, 2007: 045821. https://doi.org/10.1155/2007/45821
[8] Manohar, K., Rao, P. (2006). Speech enhancement in nonstationary noise environments using noise properties. Speech Communication, 48(1): 96-109. https://doi.org/10.1016/j.specom.2005.08.002
[9] Sigg, C.D., Dikk, T., Buhmann, J.M. (2012). Speech enhancement using generative dictionary learning. IEEE Transactions on Audio, Speech, and Language Processing, 20(6): 1698-1712. https://doi.org/10.1109/TASL.2012.2187194
[10] Abolhassani, A.H., Selouani, S.A., O'Shaughnessy, D., Harkat, M.F. (2007) Speech enhancement using PCA and variance of the reconstruction error model identification. Proc. Interspeech, pp. 974-977. https://doi.org/10.21437/Interspeech.2007-346
[11] Balcan, D.C., Rosca, J. (2006). Independent component analysis for speech enhancement with missing TF content. In: Rosca, J., Erdogmus, D., Príncipe, J.C., Haykin, S. (eds) Independent Component Analysis and Blind Signal Separation. ICA 2006. Lecture Notes in Computer Science, vol 3889. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11679363_69
[12] Guo, X., Jia, H., Wang, D. (2016). Speech enhancement using the improved k-SVD algorithm by subspace. Journal of Xidian University (Natural Science Edition), 43(6).
[13] Wilson, K.W., Raj, B., Smaragdis, R. (2008). Regularized Non-negative matrix factorization with temporal dependencies for speech denoising. Proc. Interspeech 2008, pp. 411-414. https://doi.org/10.21437/Interspeech.2008-49
[14] Duong, V.H., Bui, M.Q., Wang, J.C. (2019). Dictionary learning-Based speech Enhancement. Active Learning – Beyond Future, Book Chapter 6. https://doi.org/10.5772/intechopen.85308
[15] Candès, E.J., Li, X., Ma, Y., Wright, J. (2011). Robust principal component analysis. Journal of the ACM (JACM), 58(3): 1-37. https://doi.org/10.1145/1970392.1970395
[16] Wright, J., Ganesh, A., Rao, S., Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. Advances in Neural Information Processing Systems, pp. 2080-2088. http://hdl.handle.net/2142/74349.
[17] Gillis, N., Glineur, F. (2011). Low-rank matrix approximation with weights or missing data is np-hard. SIAM Journal on Matrix Analysis and Applications, 32(4): 1149-1165. https://doi.org/10.1137/110820361
[18] ITU Recommendation ITU-R. BS. 1387-1 method for objective measurements of perceived audio quality, tech. rep., https://www.itu.int/rec/R-REC-BS.1387, accessed on 16 February 2022.
[19] Vincent, E., Jafari, M.G., Plumbley, M.D. (2006). Preliminary guidelines for subjective evaluation of audio source separation algorithms. ICA Research Network International Workshop, Liverpool, UK, pp. 93-96.
[20] Hu, Y., Loizou, P.C, (2008). Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 16(1): 229-238. https://doi.org/10.1109/TASL.2007.911054
[21] Gu, S., Xie, Q., Meng, D., Zuo, W., Feng, X., Zhang, L. (2017). Weighted nuclear norm minimization, and its applications to low-level vision. International Journal of Computer Vision, 121(2): 183-208. https://doi.org/10.1007/s11263-016-0930-5
[22] Cai, J.F., Candès, E.J., Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4): 1956-1982. https://doi.org/10.1137/080738970
[23] Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1): 1-122. http://dx.doi.org/10.1561/2200000016
[24] Huang, P.S., Chen, S.D., Smaragdis, P., Hasegawa-Johnson, M. (2012). Singing voice separation from monaural recordings using robust principal component analysis. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 57-60. https://doi.org/10.1109/ICASSP.2012.6287816
[25] Sorensen, K.V., Andersen, S.V. (2005). Speech enhancement with natural-sounding residual noise based on connected time-frequency speech presence regions. EURASIP Journal on Advances in Signal Processing, 2005(18): 305909. https://doi.org/10.1155/ASP.2005.2954
[26] Rangachari, S., Loizou, P.C. (2006). A noise-estimation algorithm for highly non-stationary environments. Speech Communication, 48(2): 220-231. https://doi.org/10.1016/j.specom.2005.08.005
[27] Zhou, T., Tao, D. (2011). GoDec randomized low-rank & sparse matrix decomposition in noisy case. Proceedings of the 28th International Conference on International Conference on Machine Learning, pp 33-40.
[28] Gu, S., Zhang, L., Zuo, W., Feng, X. (2014). Weighted nuclear norm minimization with application to image denoising. 2014 IEEE Conference on Computer Vision

and Pattern Recognition, Columbus, pp. 2862-2869. https://doi.org/10.1109/cvpr.2014.366

[29] Xie, Y., Gu., S., Liu., Y., Zuo, W., Zhang, W. (2016). Weighted schatten p-norm minimization for image denoising and background subtraction. IEEE Transactions on Image Processing, 25(10): 4842-4857. https://doi.org/10.1109/TIP.2016.2599290

[30] Hu, Y., Loizou, P.C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. Speech Communication, 49(7-8): 588-601. https://doi.org/10.1016/j.specom.2006.12.006

[31] Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V. (1993). TIMIT acoustic-phonetic continuous speech corpus. LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/17gk-bn40

[32] Panayotov, V., Chen, G., Povey, D., Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964

[33] Daniel, P., Arnab, G., Gilles, B. (2011). The Kaldi speech recognition toolkit, speech recognition project. https://github.com/kaldi-asr/kaldi.git/Kaldi.

[34] Mavaddaty, S., Ahadi, S.M., Seyedin, S. (2016). A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation. Speech Communication, 76: 42-60. https://doi.org/10.1016/j.specom.2015.11.003

[35] Sun, P., Qin, J. (2016). Low rank and sparsity analysis applied to speech enhancement via online estimated dictionary. IEEE Signal Processing Letters, 23(12): 1862-1866. https://doi.org/10.1109/LSP.2016.2627029

[36] Huang, J., Zhang, X., Zhang, Y., Zou, X., Zeng, L. (2014). Speech denoising via low-rank and sparse matrix decomposition. ETRI Journal, 36(1): 167-170. https://doi.org/10.4218/etrij.14.0213.0033

[37] Rix, A.W., Beerends, J.G., Holler, M.P., Hekstra, A.P. (2001). Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), pp. 749-752. https://doi.org/10.1109/ICASSP.2001.941023

[38] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio Speech and Language Processing, 19(7): 2125-2136. https://doi.org/10.1109/TASL.2011.2114881

[39] Saleem, N. (2016). Single channel noise reduction system in low SNR. International Journal of Speech Technology, 20: 89-90. https://doi.org/10.1007/s10772-016-9391-z

[40] Saleem, N., Ijaz, G. (2018). Low rank sparse decomposition model based speech enhancement using gammatone filterbank and Kullback–Leibler divergence. International Journal of Speech Technology, 21: 217-231. https://doi.org/10.1007/s10772-018-9500-2