Low Rank and Sparsity Analysis Applied to Speech Enhancement via Online Estimated Dictionary

Pengfei Sun  
Southern Illinois University Carbondale

Jun Qin  
Southern Illinois University Carbondale, jqin@siu.edu

Follow this and additional works at: http://opensiuc.lib.siu.edu/ece_articles
(c) 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other users, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works for resale or redistribution to servers or lists, or reuse of any copyrighted components of this work in other works.

Recommended Citation
Low Rank and Sparsity Analysis Applied to Speech Enhancement via Online Estimated Dictionary

Pengfei Sun, Student Member, and Jun Qin, Member, IEEE

Abstract—In this letter, we propose an online estimated dictionary based single-channel speech enhancement algorithm, which focuses on low-rank and sparse matrix decomposition. In the proposed algorithm, a noisy speech spectrogram can be decomposed into low rank background noise components and an activation of the online speech dictionary, on which both low-rank and sparsity constraints are imposed. This decomposition takes the advantage of local estimated exemplar’s high expressiveness on speech components and also accommodates nonstationary background noise. The local dictionary can be obtained through estimating the speech presence probability (SPP) by applying Expectation-Maximal algorithm, in which a generalized Gamma prior for speech magnitude spectrum is used. The proposed algorithm is evaluated using signal-to-distortion ratio (SDR), and perceptual evaluation of speech quality (PESQ). The results show that the proposed algorithm achieves significant improvements at various SNRs when compared to four other speech enhancement algorithms, including improved Karhunen-Loeve transform (KLT) approach, SPP based MMSE (MMSE-SPP), NMF based RPCA (NMF-RPCA), and RPCA.

Index Terms—speech enhancement, online speech dictionary, low rank, sparsity, speech presence probability.

I. INTRODUCTION

SINGLE channel speech enhancement is a key issue in speech processing, aiming to improve the performance of speech communication systems in noisy environments. Recently-developed robust principal component analysis (RPCA) has been shown effective in separating the speech components from background noise [4], [14], [16]. This approach decomposes the spectrogram matrix as the sum of a sparse matrix and a low-rank matrix, representing speech and noise, respectively [15]. Based on the observation that unpredictable background noise is often less spectrally diverse than the foreground speech, it regards that sparse matrix represents the speech components while noise is included in the low rank matrix [1]. Because the activation of the low-rank components can be temporally variable, this unsupervised decomposition can accommodate nonstationary noise [13]. However, the targeted speech may also contain low rank components described by limited spectral bases, and these low rank speech components can be wrongly decomposed into noise matrix. Thus unsupervised decomposition may not provide accurate separation of foreground and background in certain circumstance, e.g., nonstationary noise [11].

To avoid simply using sparsity to characterize speech spectrogram, incorporating the knowledge about the likely form of the targeted speech has been introduced [2], [11]. Using nonnegative matrix factorization (NMF) technique, pre-learned explicit speech dictionary [3], [12] based on large dataset can improve the performance of RPCA. The advantage of the offline-trained global dictionary based RPCA approach is its ability to maintain the flexibility to distinguish speech and noise, and also avoid wrongly decomposing the low-rank speech components into noise subspace. However, offline-learned global basis spectra may either lead to a non-sparse activation matrix or wrongly interpret background noise as speech. This is intrinsically due to overfitting caused by limited local speech eigenvectors and shared basis spectra between speech and noise [3]. Recent study [10] indicates that exemplar based dictionary can be more effective on covering speech spectra convex hull. In addition, online trained dictionary [3] presents a solution to achieve higher noise reduction whereas lower speech distortion. To exploit the advantages of both, a natural idea is to develop an online learned exemplar-structure dictionary.

In this study, we propose an online estimated dictionary based low rank and sparse decomposition (LDLSD) algorithm. The developed local dictionary estimation inherits the merit of both exemplar’s efficient explanation on the speech manifold and adaptive speech subspace that alleviating the immediate speech distortion [4]. In LDLSD algorithm, a semi-supervised Expectation-Maximum (EM) algorithm based on a generalized Gamma speech distribution model is used to calculate the speech presence probability (SPP), and further obtain the local estimated exemplar. Moreover, the activation matrix of speech is optimized with respect to both low rank and sparsity constraints to suppress the noise entries mixed in the exemplar.

II. PROPOSED SPEECH ENHANCEMENT METHOD

A. Model

RPCA related approaches generally decompose an input spectrogram matrix $Y \in \mathbb{R}^{N \times M}$ as the summation of a low-rank matrix $L$ reflecting less spectrally diverse noise, and a sparse matrix $S$, representing the sparsity of speech energy as shown in (1). Outlying entries $E$ are frequently added to provide a relaxed estimation on the noise residuals [1], [14].

$$\min_{S, L} \|S\|_1 + \gamma_1 \|L\|_* + \gamma_2 \|E\|^2_F$$

s.t. $Y = S + L + E$

$\| \cdot \|_1$ is the nuclear norm of the matrix. The sparsity of $S$ is measured by the $L_1$ norm $\| \cdot \|_1$, and the Frobenious
norm is used to represent the noise residual. Based on the observation that speech spectrogram is sparsity in short time period but with repeated structures (i.e., low-rank) within several consecutive segments, \( S \) in (1) can be replaced by \( Y_1 S \), where as an explicit dictionary of speech spectral templates, \( Y_1 \) is multiplied by a set of temporal activation \( S \). Generally, \( Y_1 \) is a global offline trained dictionary [2]. However, such kind of final learned dictionary may lose focus on the intermediate estimation of the local speech, and accordingly, is not necessarily good enough to explain current speech frame. Instead, an online speech exemplar can be estimated based on SPP given by

\[
Y_1 = Y \odot \mathcal{P}
\]

(2)

\( \odot \) refers to element-wise product of matrices, and \( \mathcal{P} \) is the SPP, for each frequency bin defined as

\[
\mathcal{P}_j = p(z_1 \mid y_j, \lambda)
\]

\[
= \frac{p(y_j \mid z_1, \lambda)p(z_1)}{\sum_{n=0}^{n-1} p(y_j \mid z_n, \lambda)p(z_n)} = \frac{p(y_j \mid z_1, \lambda)w_1}{\sum_{n=0}^{n-1} p(y_j \mid z_n, \lambda)w_n}
\]

(3)

where index \( n \in \{1, 0\} \) represents the case of speech present and absent, respectively, \( \mu_n \) is the corresponding a priori probability that \( (\sum_n w_n = 1) \). \( \lambda \triangleq \{\mu_n, \sigma_n, w_n\} \) is the parameter set (i.e., mean, variance, and priori) for the statistic models of speech and noise. Considering noise obeys complex Gaussian distribution, \( p(y_j \mid z_0, \lambda) \) should be Rayleigh distribution, and specifically, \( p(y_j \mid z_1, \lambda) \) is assumed as generalized Gamma distribution given by [6]

\[
p(y_j \mid z_0, \lambda) = \frac{y_j^{\beta} \exp\left(-y_j^2/2\sigma^2\right)}{\delta^2} \tag{4}
\]

\[
p(y_j \mid z_1, \lambda) = \frac{\beta\nu}{\Gamma(\nu)}y_j^{\nu-1}\exp\left(-\beta y_j\right) \tag{5}
\]

where

\[
\delta = \mu_0\sqrt{2/\pi}; \quad \beta = \mu_1/\sigma_1^2; \quad \nu = \mu_2^2/\sigma_2^4 \tag{6}
\]

Usually, the local exemplar \( Y_1 \) promises a sparse activation matrix \( S \) as the representation of individual speech segment from the same subspace. However, in low SNR scenario, the estimated local exemplar will be highly corrupted by noise components, and such kind of speech 'bases' is often overcomplete. Hence there can be many feasible solutions to \( Y = Y_1 S + L + E \). To address this issue, both the most sparse and lowest rank criteria are imposed on \( S \). The main consideration is that in the estimated dictionary, speech components have larger or at least comparable magnitudes than noise entries, and the low-rank constraint imposed on \( S \) obviously utilizes the similar frequency structures in speech spectral bases. Comparatively, the residual noises in the estimated dictionary can be regarded as sparse components. Thus, we seek a representation \( S \) by solving the following optimization problem

\[
\min_{S, L} \quad \|S\|_\ast + \beta\|A\|_1 + \gamma_1\|L\|_\ast + \gamma_c\|E\|_{2, 1}
\]

\[
s.t. \quad Y = Y_1 S + L + E, S = A.
\]

(7)

where an auxiliary variable \( A \) is introduced to make the objective function separable. \( \|E\|_{2, 1} = \sum_{i=1}^{M} \sqrt{\sum_{j=1}^{N} E_{ij}^2} \) is called the \( \ell_{2, 1} \) norm. The \( \ell_{2, 1} \) norm encourages the columns of \( E \) to be zero, which assumes that the outlying entries are "sample-specific".

B. Algorithm

For each row of spectrogram matrix, the probability density function (PDF) is given by

\[
p(y | \lambda) = \prod_{j=1}^{M} p(y_j | \lambda) \tag{8}
\]

where \( p(y_j | \lambda) = \sum_{z_n} p(y_j \mid z_n, \lambda)p(z_n) \). The parameter set \( \lambda \) is estimated by maximizing the above PDF function. The following are the typical EM re-estimation formulas

\[
\hat{w}_{0,n} = \frac{1}{M} \sum_{j=1}^{M} p(z_n \mid y_j, \lambda_0) \tag{9}
\]

\[
\hat{\mu}_{0,n} = \frac{\sum_{j=1}^{M} y_j p(z_n \mid y_j, \lambda_0)}{M \hat{w}_{0,n}} \tag{10}
\]

\[
\hat{\sigma}^2_{0,n} = \frac{\sum_{j=1}^{M} (y_j - \hat{\mu}_{0,n})^2 p(z_n \mid y_j, \lambda_0)}{M \hat{w}_{0,n}} \tag{11}
\]

where \( p(z_n \mid y_j, \lambda_0) \) is obtained through (3)-(6) with the old parameter set \( \lambda_0 \). \( \lambda_0 \sim \{\hat{w}_{0,n}, \hat{\mu}_{0,n}, \hat{\sigma}^2_{0,n}\} \) denotes the new parameter set re-estimated from \( \lambda_0 \). In the next iteration, \( \lambda_0 \) is replaced by \( \lambda_0 \). The initial parameters are obtained by K-means, and the speech model parameter \( \nu \) is fixed if the mean values of speech and noise components are too close. This iteration continues until EM algorithm converges. Based on the obtained local SPP, an online updated schematic is proposed [17] as described in Algorithm 1.

Comparing with conventional Gaussian speech model, the generalized Gamma distribution is more restrictive on noise basis spectra and can accurately pick up the speech components [5]. In addition, the online update scheme allows the parameter set to be highly describable on local distribution. To solve (7), a recently developed method called the linearized alternating direction method with adaptive penalty (LADMAP) has been applied to obtain the optimization result [18]. The augmented Lagrangian function of (7) is

\[
\mathcal{L}(S, A, L, E, \rho, \Delta_1, \Delta_2) = \|S\|_\ast + \beta\|A\|_1 + \gamma_1\|L\|_\ast + \gamma_c\|E\|_{2, 1} + \frac{\rho}{2}\|Y - Y_1 S - L - E\|_F^2 + \frac{\Delta_1}{\rho}\|S - A\|_F^2 + \frac{\Delta_2}{\rho}\|E\|_F^2 \tag{12}
\]

where \( \Delta_1 \) and \( \Delta_2 \) are Lagrangian multipliers. With some algebra, the updating schemes are outlined in Algorithm 2. \( \Theta, \Theta_S, \text{ and } \Omega \) are the singular value thresholding, shrinkage, and the \( \ell_{2, 1} \) minimization operator, respectively, and \( \eta = \|Y_1\|_2^2 \) [18].
Algorithm 1 Online estimated SPP by EM algorithm

Input: speech spectrogram matrix \( Y \in \mathbb{R}^{N \times M} \)

Initialize: Set maxIter, tolerance \( \epsilon \), threshold \( \theta_i \), where Step1-3 is conducted in \( i \)th row, \( i \leq N \).

Step 1 Pre-classification:
1. Initialize \( \lambda_0 = \{ \mu_{0,n}, \sigma_{0,n}, \nu_{0,n} \} \) by K-means.
2. If \( ||\mu_{0,n} - \mu_{0,0}|| \leq \theta_i \) then
3. Fixed \( \nu_i \).
4. End if

Step 2 EM based SPP:
5. While \( ||\lambda_0 - \lambda_0|| \geq \epsilon \) or (k ≤ maxIter) do
6. E-step: Calculate the expectation PDF in (8) by (3)-(6).
7. M-Step: Calculate the new parameter set \( \lambda_0 = \{ \mu_{0,n}, \nu_{0,n}, \sigma_{0,n} \} \) by (9)-(11), and \( \lambda_0 \leftarrow \lambda_0 \).
8. End while

Step 3 Online updated SPP:
9. Input: \( \lambda_0 \rightarrow \lambda_j (j \leq M) \) and parameter \( \alpha \).
10. Update \( \Lambda \) for each new input \( y_j (j > M) \):

\[
\begin{align*}
\mu_{j,n} &= \alpha (w_{j,n} - \mu_{j,n}) + (1 - \alpha)p(z_n | y_j, \lambda_{j-1}) \\
\sigma^2_{j,n} &= \alpha (\frac{w_{j,n} - \mu_{j,n}}{w_{j,n}})^2 + (1 - \alpha)p(z_n | y_j, \lambda_{j-1})^2 \\
\end{align*}
\]

Output: SPP matrix \( \mathcal{P}_{ij} = p \{ z_1 | y_i, \lambda \} \)

### III. EXPERIMENTAL EVALUATION

The noisy speech signals were synthesized by adding speech samples to different types of noises at various input SNRs (i.e., -10, -5, 0, 5, and 10 dB). Thirty speech samples were selected from NOIZEUS database, and 30 were randomly selected from IEEE wide band speech dataset [9]. Nine different noise samples were used, including six noises (i.e., car, babble, airport, exhibition, restaurant, and train) from AURORA database, two simulated noises (i.e., Gaussian and pink noise), and one jackhammer noise sample from [9]. All signals were resampled to 8 kHz sampling rate, and the spectrograms were calculated with a window length of 32 ms, and a hop of 10 ms. The performance of the proposed LDLSD algorithm was evaluated by comparing with four other algorithms, including one conventional subspace approach (i.e., KLT [7]), and three state-of-the-arts (i.e., MMSE-SPP [5], NMF-RPCA [2], and RPCA [8]). In NMF-RPCA algorithm, the speech dictionary \( Y_1 \) was learned from the spectrograms of all 60 speech utterances used in this study. Sparse NMF with a generalized KL-divergence [2] was used to obtain the dictionary, which consisted of 300 bases. In other words, 5 basis vectors were extracted for each speech utterance.

An intuitive comparison of the improved speech spectrogram by the proposed LDLSD and two RPCA based algorithms has been shown in Fig.1. For NMF-RPCA, the artificial frequency components can be found around 2.7 second (as circled in Fig.1c). It indicates that the global dictionary bases may lead to an overfitting situation. In addition, the overlap of speech and noise basis convex hull can cause speech-similar-structure noise components (as circled around 1 second in Fig.1c). For RPCA, low rank speech spectrum ingredients are very likely to be wrongly decomposed into noise subspace (as circled at the frequency band 0.5-1 kHz in Fig.1f). Comparatively, the proposed LDLSD demonstrates better decomposition results shown in Fig.1g and 1h. The majority of noise components are correctly decomposed into \( L \). Moreover, the speech matrix \( Y_1, 5 \) in Fig.1g includes most low frequency components, and has least signal distortions than the speech matrix obtained by NMF-RPCA and RPCA algorithms in Fig.1c and 1e.

Two metrics, signal-to-distortion ratio (SDR) calculated by BSS_EVAL package and perceptual evaluation of speech quality (PESQ), are used to evaluate speech enhancement algorithms. Figure 2a shows that the averaged SDRs of the enhanced speeches by applying five different algorithms. The proposed LDLSD algorithm demonstrates the highest SDRs at all SNRs (-10, -5, 0, 5, and 10 dB). It indicates that the proposed algorithm can more effectively separate speech components from background noises. In addition, compared with NMF-RPCA, the LDLSD has a comparable averaged SDR at \( SNR = -10 \) dB, but demonstrates significantly higher averaged SDRs than the NMF-RPCA at other SNRs (i.e., -5, 0, 5, and 10 dB).

The PESQ scores of enhanced speeches by five algorithms are shown in Fig.2b. The LDLSD algorithm shows signifi-
and 0.4 higher than KLT. Especially at low SNRs (-10 and -5
dB), the low rank and sparse criteria imposed on the activation
matrix $S$ help achieve a better performance than NMF-RPCA
algorithm.

In addition, the jackhammer noise, as a highly transient
noise, is applied to evaluate the performance of the proposed
algorithm. Figure 3 shows the SDRs and PESQ of the en-
anced speeches from a jackhammer noise background at
various SNRs by five algorithms. The LDLSD demonstrates
an obvious advantage over other four algorithms on both
two metrics. In magnitude spectra space, the transient noises
(e.g., jackhammer noises) and global speech dictionary may
be partially overlapped, for example the impulsive components
are quite similar to the speech fricatives. This can cause an
ambiguity in speech and noise separation in NMF-RPCA.
Comparatively, the proposed LDLSD algorithm has two obvi-
ous merits: 1) the local estimated exemplar can help to exclude
most of the transient features; 2) the low rank constraint
imposed on the activation matrix $S$ can also reduce the impact
of transient noise residuals in the online estimated dictionary
$Y_1$.

IV. CONCLUSION

In this letter, we investigate how the SPP based local
speech dictionary can be employed in the subspace framework
to obtain the low rank and sparse components of speech
spectrogram for noisy speech enhancement. A local dictionary
based low rank and sparsity decomposition has been proposed
to separate the noise and speech components. An online
updated EM algorithm is introduced to obtain SPP matrix
according to the input noisy speech matrix. By multiplying
this SPP matrix element-wise, the broad bases in the speech
subspace can be reduced, which consequently improves the
accuracy of local speech dictionary. Moreover, the online
estimated dictionary is sufficient enough in basis subspace
to avoid speech distortion. Specifically, the most sparsity and
lowest rank criteria are both imposed to the activation matrix
to achieve a noise-resistant decomposition. The results show
that LDLSD algorithm obtains significant improvements at
various SNRs w.r.t SDR and PESQ, compared with four
algorithms, including KLT, MMSE-SPP, NMF-RPACR and
PCA. The future work of this study includes investigation on
noise constraints, such as noise variance and noise modeling.

REFERENCES

wavelet packet of an improved principal component analysis. Computer


