SPEECH ENHANCEMENT BY SPARSE, LOW-RANK, AND DICTIONARY SPECTROGRAM DECOMPOSITION

Zhuo Chen

LabROSA,
Columbia University,
zc2204@columbia.edu

Daniel P.W. Ellis

LabROSA,
Columbia University,
dpwe@ee.columbia.edu

ABSTRACT

Speech enhancement requires some principle by which to distinguish speech and noise, and the most successful separation requires strong models for both speech and noise. If, however, the noise encountered differs significantly from the system's assumptions, performance will suffer. In this work, we propose a novel speech enhancement system based on decomposing the spectrogram into sparse activation of a dictionary of target speech templates, and a low-rank background model, which makes few assumptions about the noise other than its limited spectral variation. A variation of this model specifically designed to handle transient noise intrusions is also proposed. Evaluation via BSS_EVAL and PESQ show that the new approaches improve signal-to-distortion ratio in most cases and PESQ in high-noise conditions when compared to several traditional speech enhancement algorithms including log-MMSE.

Index Terms— speech enhancement, spectrogram decomposition, sparse, low-rank, robust PCA

1. INTRODUCTION

Enhancing degraded and noisy recordings of speech is a key problem in speech processing, as a preprocessor both for automatic speech recognition and for human listeners. Since noise is usually unpredictable and highly variable, it can be difficult to formulate constraints on the noise components that are both adequately specific to support good-quality separation, and yet sufficiently broad to handle unseen noise. Existing speech enhancement systems make a number of assumptions about the noise, including stationarity and/or low magnitude [1], or directly pre-fix the spectra [2] or the rank of the noise [3]. When the actual noise fails to match these assumptions, enhancement rapidly declines. The high unpredictability of noisy interference means that speech enhancement performance cannot be guaranteed for real-world applications.

The recently-developed technique of Robust Principal Component Analysis (RPCA) [4] provides a new approach to distinguishing the background noise. RPCA decomposes a matrix into two parts, "sparse" and "low-rank", based on a well-behaved convex optimization. RPCA has recently been successfully applied to the separation of foreground singing and background music [5], indicating the technique's ability to distinguish a more regular background from a more variable foreground. In a speech enhancement scenario, even unpredictable background noise is often less spectrally diverse than the foreground speech, indicating that RPCA could be beneficial.

In this work, we further decompose the sparse component from RPCA into the product of a pre-learned dictionary of spectra with a sparse activation matrix. Based on this decomposition we propose a novel speech enhancement algorithm to identify background noise as the sum of a low-rank matrix and a residual. Unlike other approaches, the only assumption about the noise component is that its spectrogram can be accurately modeled as a low rank part and residual, where the low-rank "basis" adapts to particular signal at hand, making it better able to deal with unseen noise. Because the activation of the low-rank component can be temporally variable, this approach can accommodate not only stationary noise, but also many transient noises, provided their spectral variability is limited. Moreover, because the fitting of the speech and noise models is accomplished by simultaneous optimization, the rank of the noise model will adapt to the conditions in each individual utterance, greatly reducing the problem of mixing between separated speech and noise that plagues many other dictionary-based systems.

To handle spectrally diverse transients, a variation of the decomposition can be made to minimize the L_1-norm of the residual instead of the L_2-norm. In the right conditions, this will give the system a better ability to deal with the large but sparse noise, thereby improving performance.

In the next section, we introduce the model and algorithm in more detail. Section 3 will describe experiments, and the results are discussed in section 4. Finally, we draw conclusions in section 5.

2. THE PROPOSED MODEL

2.1. Model

Through RPCA, an input matrix can be decomposed as the summation of a low rank matrix and a sparse matrix. The basic RPCA model is shown in eqn (1), in which Y is the input matrix, S refers to sparse part, and L refers to the low-rank part. $\|\cdot\|_*$ refers to the nuclear norm of the matrix, which is the summation of its singular values, which is a proxy for minimizing the rank of the matrix L. The sparsity of S is measured by the L_1-norm $\|\cdot\|_1$, meaning the

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0014. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

summation of absolute value of the matrix's elements.

$$\begin{aligned} &\min_{S,L} \|S\|_1 + \lambda \|L\|_* \\ &\text{s.t. } Y = S + L \end{aligned} \tag{1}$$

When Y is a spectrogram, a low-rank matrix is one in which all the time-columns can be spanned by only a few spectral basis slices, and a sparse matrix corresponds with sparsity of signal energy in a short-time Fourier transform decomposition.

It is both a strength and a weakness of RPCA that it is totally unsupervised: the distinction between sparse and low-rank components is based on the particular properties of each individual sample, and can give widely varying results in response to small changes in the statistics of the component signals. In the speech enhancement application, we may expect certain kinds of noise to be low-rank, but the target speech may also be described by a limited number of spectral bases, thus it is not immediately obvious how we would expect speech and noise to be distributed among the sparse and low-rank components. By the same token, incorporating this knowledge about the likely form of the target speech into the separation scheme should allow us to improve system performance. This can be accomplished by replacing the sparse matrix S in eqn. (1) with an explicit, fixed dictionary of speech spectral templates, W, multiplied by a set of sparse temporal activations H to give:

$$\begin{aligned} &\min_{H,L} \|H\|_1 + \lambda \|L\|_* \\ &\text{s.t. } Y = WH + L \end{aligned} \tag{2}$$

For speech enhancement, we can interpret this decomposition as follows: Background noise with little variation in spectrum (even if it has substantial variation in amplitude) will be successfully captured by the low-rank component L. Conversely, the fixed set of spectral bases, W, combine with their sparse activations, H, to form a product that is constrained to consist of speech-like spectra.

Because an energy distribution such as a spectrogram is intrinsically non-negative, we can improve the suitability of the model by extending it to include nonnegativity constraints on the activations H. Also, since there will likely be a low level of full-rank random variation in the spectral columns, it can help to include a conventional mean-squared error (i.e., Gaussian noise) term in the decomposition, leading to:

$$\begin{aligned} &\min_{H,L,E} \frac{1}{2}\|E\|_2^2 + \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \mathcal{I}_+(H) \\ &\text{s.t. } Y = WH + L + E \end{aligned} \tag{3}$$

where $\mathcal{I}_+(H)$ is the auxiliary function to provide the nonnegativity constraints, which has value of infinity where H is negative and has zero elsewhere, E is the Gaussian noise residual, and λ_L and λ_H are two weighting terms to control the optimization.

Note that without L this model would be equivalent to sparse NMF, and speech enhancement approaches along these lines have been previously proposed [3]. Our formulation, however, has several advantages in comparison with NMF: To estimate the noise under plain NMF, the noise spectra – or at least the rank of the noise – must be determined explicitly in advance. But in practical situations the rank of different noise types can vary greatly – the noise of keyboard tapping will generally be much lower rank than, say, restaurant noise. If a fixed noise rank is set too low, it will be unable to capture the noise. But if it is set too high, the extra noise dimensions will start to capture parts of the target speech, leading to target energy loss. A second advantage of our approach is that separating the speech and noise components makes it easier to place the most appropriate constraints on each part. For example, it is important to impose a sparsity penalty on the speech component to avoid fitting background noise with multiple, simultaneous speech codewords. At the same time, such a constraint is usually undesirable for the noise component, where we would like to cover a wider range of variation as efficiently as possible. These differing constraints are easily applied when target speech and background noise correspond to different components, as in our model. A final advantage is that our model is more naturally extendable, for instance by replacing the target speech term with a more complex model capturing more of the intrinsic structure of speech, but the noise model can be kept as-is. The possible extensions are discussed in section 4.1.

Algorithm 1 Proposed model with L_2 residual

Input: Y, W
Output: H, L
Initialization: $H = random$; $L = 0$; $Z = 0$; $\Omega = 0$; $t = 1$
while not converged **do**
  update H:
  $H^{t+1} = (W^\top W + \rho I)^{-1}(W^\top (Y - L^t) + \rho(Z^t - \Omega^t))$
  update Z,L:
  $U\Sigma V = svd(Y - WH^{t+1})$; $L^{t+1} = U\mathcal{S}_{\lambda_L}(\Sigma)V$
  $Z^{t+1} = \mathcal{S}_{+\lambda_H/\rho}(H^{t+1} + \Omega^t)$
  update Ω:
  $\Omega^{t+1} = \Omega^t + H^{t+1} - Z^{t+1}$
  $t = t + 1$
end while

2.2. Algorithm

The model of eqn. (3) can be interpreted as the following optimization problem:

$$\min_{H,L} \frac{1}{2}\|Y - WH - L\|_2^2 + \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \mathcal{I}_+(H) \tag{4}$$

We can solve this with the Alternating Direction Method of Multipliers (ADMM) [7] by introducing an auxiliary parameter Z with an associated equality constraint:

$$\begin{aligned} &\min_{H,L} \frac{1}{2}\|Y - WH - L\|_2^2 + \lambda_H \|Z\|_1 + \lambda_L \|L\|_* + \mathcal{I}_+(Z) \\ &\text{s.t. } Z = H \end{aligned} \tag{5}$$

By introducing the scaled dual variable Ω and the scaling parameter $\rho > 0$, the augmented Langrangian function is formulated as:

$$\begin{aligned} \mathcal{L}_\rho = &\frac{1}{2}\|Y - WH - L\|_2^2 + \lambda_H \|Z\|_1 + \lambda_L \|L\|_* \\ &+ \frac{\rho}{2}\|H - Z + \Omega\|_2^2 + \mathcal{I}_+(Z) \end{aligned} \tag{6}$$

The problem can then be solved by alternately updating H, L, Z, and Ω, while holding the other parameters fixed, as detailed in Algorithm 1. There, $\mathcal{S}_\lambda(\cdot)$ refers to the well-known soft-threshold operator [4], and $\mathcal{S}_{+\lambda}(\cdot)$ indicates the additional non-negative projection step after the soft-threshold step.

Algorithm 2 Transient model with L_1 residual

Input: Y, W
Output: H, L, S
Initialization: $H = random$; $L = 0$; $S = 0$; $Z = 0$; $\Omega_1 = 0$; $\Omega_2 = 0$; $t = 1$
while not converged **do**
  update H:
  $M = Y - L^t - S^t + \Omega_2^t$
  $H^{t+1} = (W^\top W + \rho I)^{-1}(W^\top M + \rho(Z^t - \Omega_1^t))$
  update Z,L, S:
  $U\Sigma V = svd(Y - WH^{t+1} - S^t + \Omega_2^t)$; $L^{t+1} = U\mathcal{S}_{\lambda_L}(\Sigma)V$
  $Z^{t+1} = \mathcal{S}_{+\lambda_H/\rho}(H^{t+1} + \Omega_1^t)$
  $S^{t+1} = \mathcal{S}_{\lambda_S}(Y - WH^{t+1} - L^{t+1} + \Omega_2^t)$
  update Ω_1, Ω_2:
  $\Omega_1^{t+1} = \Omega_1^t + H^{t+1} - Z^{t+1}$
  $\Omega_2^{t+1} = \Omega_2^t + Y - WH^{t+1} - L^{t+1} - S^{t+1}$
  $t = t + 1$
end while

2.3. Handling Transient Noise

Transient noise, meaning noise with sparse, short-duration bursts such as keyboard tapping or birds chirping, can be a particular problem for speech enhancement systems. For instance, systems that attempt to estimate a "noise floor" during the most quiet parts of the signal will entirely fail to recognize transient noise bursts. Depending on the spectral variability of the transients, they could be a problem for our approach too, so we propose a variant of the model better suited to this situation: Rather than using an L_2-norm to minimize the energy of the residual noise E in eqn. (3), we can use an L_1-norm as in the original RPCA formulation to provide a sparse residual term S able to absorb high-energy, short-duration bursts:

$$\begin{aligned} &\min_{H,L,S} \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \lambda_S \|S\|_1 + \mathcal{I}_+(H) \\ &\text{s.t. } Y = WH + L + S \end{aligned} \tag{7}$$

A similar ADMM algorithm can be used to solve this modified model. By introducing an additional Lagrangian variable Ω_2, the Lagrangian function can be re-written as below, whose solution is accomplished by Algorithm 2.

$$\begin{aligned} \mathcal{L}_\rho = &\frac{1}{2}\|Y - WH - L - S + \Omega_2\|_2^2 + \lambda_H \|Z\|_1 + \lambda_L \|L\|_* \\ &+ \lambda_S \|S\|_1 + \frac{\rho}{2}\|H - Z + \Omega_1\|_2^2 + \mathcal{I}_+(Z) \end{aligned} \tag{8}$$

3. EXPERIMENTS AND RESULT

The proposed systems were evaluated with 5000 noisy speech examples, totaling 3.5 hours. The noisy signals were synthesized by adding clean speech to a variety of noise signals at different SNRs. Speech examples were chosen at random from the TIMIT dataset, and the noise signals were drawn from the AURORA dataset and from other internet sources. We included eight "stationary" noises – car, exhibition, restaurant, babble, casino, train, subway, and airport – and four "transient" noise types – keyboard, machine gun, birds, and eating chips. We mixed the signals at five different signal-to-noise ratios (SNRs) from -10 to 10 dB. All files were resampled to 8 kHz sampling rate and pre-emphasized. To calculate the spectrograms we used a window length of 32 ms (256 points), and a hop of 10 ms (80 points). The speech dictionary W was learned from 20 speakers (200 utterances) from the TIMIT dataset, disjoint from the speakers used to make the noisy examples. Sparse NMF with a generalized KL-divergence [6] was used to create the dictionary, which consisted of 1000 bases.

We evaluated speech enhancement using two metrics. We used the popular BSS_EVAL [8] package to calculate the widely-reported Signal-to-Distortion Ratio (SDR). The second criteria was the PESQ estimate of subjective speech quality [9]. For both metrics, a larger score indicates better performance.

Four classical algorithms were compared with the proposed approaches: log-MMSE estimation [12], spectral subtraction [11], a subspace (KLT) algorithm [13], and Wiener filtering [10]. We used Loizou's implementations of these systems [1]. Tables 1 and 2 give the SDR and PESQ results, respectively, over all noise conditions, whereas tables 3 and 4 report results only for the four transient noise types, to highlight the differences on this kind of material. For comparison, the tables also give results for the sparse component obtained by unmodified RPCA (eqn. (1)). Figure 1 shows spectrograms of a single sound example and its decompositions.

	orig	L2	L1	LS	mse	ssub	klt	wn
-10dB	0	-3.6	**-3.4**	-6.8	-7.6	-8.3	-7.8	-19
-5dB	0	**1.5**	1.3	-1.5	-2.2	-1.7	-2.2	-16
0dB	0	**5.9**	4.7	3.6	3.8	3.4	2.4	-2.1
5dB	0	**10.1**	7.7	7.5	7.3	7.6	7.4	-1.8
10dB	0	**12.8**	9.8	10.2	12.7	10.2	12.8	0.5

Table 1: SDR values (in dB) for all systems at various SNRs, averaged across all noise types. WHLE is the proposed system using L_2-norm residual, WHLS uses L_1-norm, LS corresponds to the sparse component of unmodified RPCA, and mmse, ssub, KLT, and wien are the four comparison algorithms.

	WHLE	WHLS	LS	mmse	ssub	KLT	wien
-10dB	**1.0**	**1.0**	**1.0**	**1.0**	0.9	0.5	0.2
-5dB	**1.5**	**1.5**	1.1	**1.5**	**1.5**	1.1	0.7
0dB	1.7	1.8	1.3	**1.9**	**1.9**	1.6	1.3
5dB	2.0	2.2	1.5	2.3	**2.3**	2.1	1.8
10dB	2.2	2.5	1.7	**2.6**	2.5	2.5	2.2

Table 2: PESQ scores averaged across all noises. Legend as table 3.

4. DISCUSSION

From the tables, we can see that the proposed models outperformed the comparison systems in terms of SDR for all conditions, indicat-

	WHLE	WHLS	LS	mmse	ssub	KLT	wien
-10dB	-1.9	**-1.5**	-6.7	-9.6	-8.6	-8.6	-21
-5dB	**2.6**	2.1	-1.9	-5.4	-2.4	-4.1	-13
0dB	**6.3**	4.9	2.8	2.6	1.8	-0.7	-3.0
5dB	**10.3**	7.5	6.7	5.1	6.3	5.1	-4.7
10dB	**12.8**	11.6	9.6	11.3	8.8	11.53	0.2

Table 3: SDR values in dB for transient noise types (keyboard, machine gun, birds, eating chips) only.

	WHLE	WHLS	LS	mmse	ssub	KLT	wien
-10dB	1.0	**1.1**	0.9	0.9	0.9	0.9	0.2
-5dB	1.3	**1.4**	1.1	1.3	1.4	0.9	0.6
0dB	1.5	**1.7**	1.3	1.6	1.7	1.5	1.3
5dB	1.9	**2.2**	1.6	2.1	**2.2**	1.9	1.6
10dB	2.2	2.5	1.8	2.6	2.3	**2.7**	2.1

Table 4: PESQ scores for transient noise types only.

ing the ability of the approach to separate target energy from noise. The picture given by the PESQ score is a more mixed: the proposed systems perform best only when the SNR is smaller than 0 dB (or 10dB for the WHLS model). One reason for the lower PESQ scores of the proposed systems in low-noise conditions is the limited span of the dictionary which which placea an upper bound on speech reconstruction accuracy in good conditions.

4.1. Possible extensions

As mentioned in Section 2, one important advantage of the proposed framework is that the model can be easily extended by swapping in different models for the speech and/or noise independently. There are several alternatives for the speech part of the model, including using a shift-invariant dictionary to match the speech with larger time-frequency patches that can move in frequency as well as time. These could help enforce the natural continuity of speech, and thus show improve resistance to transient noise. Another very promising extension is to adapt the dictionary as well when performing the decomposition by incorporating some additional constraints particularly designed for the speech (e.g. harmonicity, continuity, etc.), which may offer a potential solution to the limitations of dictionary-based systems mentioned above.

5. CONCLUSION

In this work, we have proposed a novel framework based on a combination Robust PCA and learned target source models to solve the speech enhancement problem. Our experiments show that the two variants of the model are both able to perform a substantial separation of speech from noise interference, even in the face of highly nonstationary transient noise types. We will continue to investigate mechanisms to improve the quality of the separated target speech.

6. REFERENCES

[1] P. Loizou, *Speech Enhancement: Theory and Practice*. Taylor and Francis, 2007.

[2] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *Proceedings of Interspeech*, 2012.

[3] G. J. M. Dennis L. Sun, "Universal speech models for speaker independent single channel source separation," in *Proc. IEEE ICASSP*, 2013.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *arXiv preprint arXiv:0912.3599*, 2009.

[5] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. IEEE ICASSP*, 2012, pp. 57–60.

[6] O. R. Schmidt M.N., "Single-channel speech separation using sparse non-negative matrix factorization," in *Proceedings of Interspeech*, 2006, p. 26142617.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, pp. 1–122, 2011.

[8] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio Speech Lang. Process*, vol. 14, pp. 1462–1469, 2006.

[9] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codes," in *Proc. IEEE ICASSP*, 2001, pp. 749–752.

[10] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE ICASSP*, 1996.

[11] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhanceing speech corrupted by colored noise," in *Proc. IEEE ICASSP*, 2002.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process*, vol. 33, pp. 443–445, 1985.

[13] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise." *IEEE Trans. Speech Audio Process*, pp. 334–341, 2003.

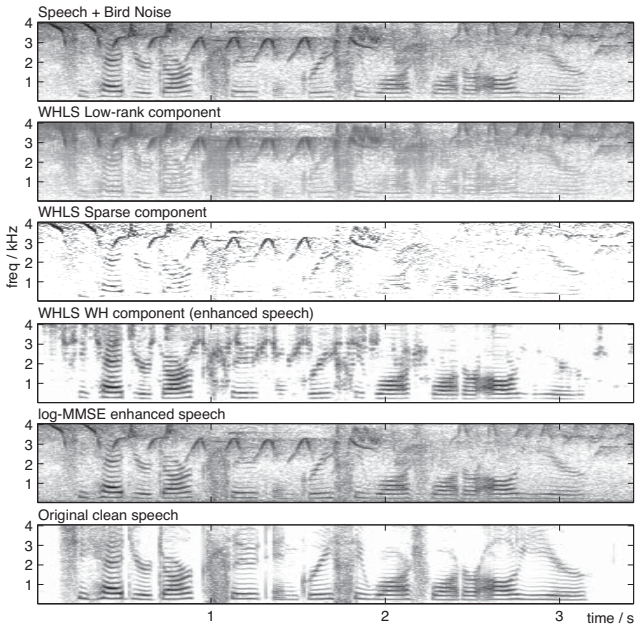

Figure 1: Example decomposition. The top pane shows the spectrogram of speech mixed with bird chirping at -5 dB. For the WHLS decomposition, the resulting low-rank and sparse parts are shown in the second and third panes, and the enhanced speech is shown in the fourth pane. For comparison, the fifth pane shows log-MMSE enhancement. The clean speech appears in the bottom pane.