Content-adaptive speech enhancement by a sparsely-activated dictionary plus low rank decomposition

Zhuo Chen, Hélène Papadopoulos, Daniel P.W. Ellis

CONTENT-ADAPTIVE SPEECH ENHANCEMENT BY A SPARSELY-ACTIVATED DICTIONARY PLUS LOW RANK DECOMPOSITION

Zhuo Chen[†] *Hélène Papadopoulos*[⋆] *Daniel P.W. Ellis* [†]

[⋆]Laboratoire des Signaux et Systèmes, UMR 8506, CNRS-SUPELEC-Univ. Paris-Sud, France
[†]LabROSA, Columbia University
zc2204[at]columbia.edu helene.papadopoulos[at]lss.supelec.fr dpwe[at]ee.columbia.edu

ABSTRACT

One powerful approach to speech enhancement employs strong models for both speech and noise, decomposing a mixture into the most likely combination. But if the noise encountered differs significantly from the system's assumptions, performance will suffer. In previous work, we proposed a speech enhancement model that decomposes the spectrogram into sparse activation of a dictionary of target speech templates, and a low-rank background model. This makes few assumptions about the noise, and gave appealing results on small excerpts of noisy speech. However, when processing whole conversations, the foreground speech may vary in its complexity and may be unevenly distributed throughout the recording, resulting in inaccurate decompositions for some segments. In this paper, we explore an adaptive formulation of our previous model that incorporates separate side information to guide the decomposition, making it able to better process entire conversations that may exhibit large variations in the speech content.

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

1. INTRODUCTION

In the context of processing and analyzing high-dimensional data, such as videos, bioinformatics or audio data, a common challenge is to extract useful information from a massive amount of related or unrelated data in a complex environment. Very often, the problem can be formulated as separating the foreground components from an underlying background as, for instance, when separating the moving objects from the stable environment in video surveillance [1]. Robust Principal Component Analysis (RPCA [2, 3]) is a technique that attempts to decompose signals into sparse and low-rank components, and has recently attracted substantial attention.

RPCA has been used extensively in the field of image processing (e.g. image segmentation [4], visual pattern correspondence [5], surveillance video processing [6], batch image alignment [7], etc.). The framework has also been considered for extracting information from audio signals. In the context of music signal processing, RPCA has been used to separate the singing voice from a background accompaniment in monaural polyphonic recordings [8, 9]. In the context of speech processing, RPCA has been used for the task of speech enhancement [10, 11]. In both cases, the foreground (signing voice/speech) and the background are separated by decomposing the Short-Time-Fourier Transform (STFT) magnitude (i.e., spectrogram) into sparse and low-rank components.

We are interested here in further exploring the framework of sparse and low-rank decompositions for speech enhancement. Enhancing degraded and noisy recordings of speech is a key problem 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 sufficiently broad to handle unseen noise. Existing speech enhancement systems make a number of assumptions about the noise, including stationarity and/or low magnitude [12], or explicitly fix the spectra [13] or the rank of the noise [14]. 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. However, in a speech enhancement scenario, even unpredictable background noise is often less spectrally diverse than the foreground speech, indicating that it could benefit from the RPCA's ability to distinguish a more regular background from a more variable foreground.

Based on the idea of using sparse and low-rank decompositions for speech enhancement, we recently proposed a model that further decomposes the sparse component of RPCA into the product of a pre-learned dictionary of spectra with a sparse activation matrix, and where the background noise is identified as the sum of a low-rank matrix and a residual [10]. 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. This model, here referred to as sparsely activated dictionary RPCA (SaD-RPCA), will serve as a baseline for the present work.

Sparse and low-rank decomposition-based approaches for speech processing have mostly been evaluated on short audio excerpts consisting in a single speaker and a single type of background noise. However, real-world scenarios, such as real-life conversations, involve in general several speakers that interact in a complex acoustic environment. In this case the background may include significant changes in its acoustic characteristics and dynamics which may rival the variation in the foreground (e.g. background music plus noise from the other conversations plus ambient noise in a restaurant conversation scenario), and hence its rank in the spectrogram representation. Further, the foreground may vary in its complexity (e.g., several persons possibly speaking together then one by one) and may be unevenly distributed throughout the record-

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. Part of this research was supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Program.

ing (e.g., entire segments without speech). The question of how to cope with the intrinsic structure of the analyzed data (e.g. the speech/non speech patterns) remains open.

In this article, we explore an adaptive version of SaD-RPCA that is able to handle such conversations with large variations in the foreground by adjusting the task through the incorporation of domain knowledge that guides the decomposition towards results that are physically and semantically meaningful. More specifically, we incorporate speech activity information as a cue to separate the speech voice from the background. The sparse (foreground) component should be denser in sections containing voice, while portions of the sparse matrix corresponding to non-speech segments should ideally be null. Thus, while the technique remains the same as in [10] at the lowest level, we consider the problem of segmenting a longer conversation into suitable pieces, and how to locally adapt the decomposition by incorporating prior information.

2. PROPOSED MODEL

2.1. RPCA via Principal Component Pursuit

In [2] , Candès *et al*. show that, under very broad conditions, a data matrix $Y \in \mathbb{R}^{m \times n}$ (in our case the spectrogram) can be exactly and uniquely decomposed into a low-rank component L and a sparse component S via a convex program called *Principal Component Pursuit* (RPCA-PCP) given by:

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

where $\lambda > 0$ is a regularization parameter that trades between the rank of L and the sparsity of S. The nuclear norm $\|\cdot\|_*$ (the sum of singular values) is used as surrogate for the rank of L [15], and the ℓ_1 norm $\|\cdot\|_1$ (the sum of absolute values of the matrix entries) is an effective surrogate for the ℓ_0 pseudo-norm (the number of non-zero entries in the matrix [16, 17]).

The Augmented Lagrange Multiplier Method (ALM) and its practical variant, the Alternating Direction Method of Multipliers (ADMM), have been proposed as efficient optimization schemes to solve this problem [18, 19, 20].

2.2. Sparsely Activated Dictionary RPCA (SaD-RPCA)

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. We have thus proposed in [10] replacing the sparse matrix S in Eq. (1) with an explicit, fixed dictionary of speech spectral templates, W, multiplied by a set of sparse temporal activations H. With this model, we expect the background noise to have little variation in spectrum even if it has substantial variation in amplitude, and thus to 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. We further improve the suitability of the model by extending it to include nonnegativity constraints on the activations H, since the spectrogram is intrinsically a non-negative energy distribution. Finally, since there will likely be a low level of full-rank random variation in the spectral columns, we include a conventional mean-squared error (i.e., Gaussian noise) term in the decomposition. This leads to the following model:

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

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 [14]. For a more detailed comparison between SaD-RPCA and sparse NMF, we refer the reader to [10].

The optimization problem in (2) is equivalent to:

$$\min_{H,L} \frac{1}{2}\|Y - WH - L\|_2^2 + \lambda^H \|H\|_1 + \lambda^L \|L\|_* + \mathcal{I}_+(H) \tag{3}$$

To make objective function (3) separable, we introduce an auxiliary parameter Z with an associated equality constraint, leading to:

$$\begin{aligned} &\min_{H,L} \frac{1}{2}\|Y - WH - L\|_2^2 + \lambda^H \|Z\|_1 + \lambda^L \|L\|_* + \mathcal{I}_+(Z) \\ &\text{s.t.} \quad Z = H \end{aligned} \tag{4}$$

By introducing the scaled dual variable Ω and the scaling parameter $\rho > 0$, we formulate the augmented Lagrangian function of (4) as:

$$\begin{aligned} \mathcal{L}_\rho = &\frac{1}{2}\|Y - WH - L\|_2^2 + \lambda^H \|Z\|_1 + \lambda^L \|L\|_* \\ &+ \frac{\rho}{2}\|H - Z + \Omega\|_2^2 + \mathcal{I}_+(Z) \end{aligned} \tag{5}$$

Problem (4) can be solved by minimizing the augmented Lagrangian function of (5). The Alternating Direction Method of Multipliers splits the minimization of (5) into smaller and easier subproblems, by sequentially updating H, L, Z, and Ω, while holding the other parameters fixed.

2.3. Content-Adaptive SaD-RPCA (CaSaD-RPCA)

As discussed in Section 1, in a conversation, the speech signal typically exhibits a clustered distribution in the time-frequency plane relating to the structure of the conversation that consists of alternating speech and non-speech (silent) segments. This structure should be reflected in the decomposition: frames belonging to speech-inactive segments should result in zero-valued columns in Z.

The balance between the sparse and low-rank contributions is set by the value of the regularization parameters λ^H and λ^L. Experiments show that the decomposition is very sensitive to the choice of these parameters, with frequently no single set of values able to achieve a satisfying separation between speech and background across a whole conversation. This is illustrated in Fig. 1, which shows one example on one utterance of the relation between the separation quality and the choice of the regularization parameter λ^H (with λ^L fixed). As we can observe, the best λ^H differs depending on whether we process the entire utterance, or restrict processing to just the speech-active parts. Because the separation for the silent part is monotonically better as λ^H increases, the difference between the optimum λ^H indicates that the global separation quality is compromised between the speech and the noise part.

We propose an adaptive variant of the SaD-RPCA algorithm, referred to as Content-Adaptive SaD-RPCA (CaSaD-RPCA), that adapts the decomposition to the content of the signal by incorporating side information about the structure of the speech signal. Specifically, speech activity information is used as a cue to adjust the regularization parameter through the entire analyzed recording during the update of the sparse component Z, and therefore better match the balance between sparse and low-rank contributions to suit to the actual semantic content. This idea is related to previous theoretical work [21, 22, 23], but to our knowledge, its application in the framework of RPCA is new.

From Eq. (5), we can see that minimization of Z reduces to:

$$Z^{k+1} = \min_Z \left\{ \lambda^H \|Z\|_1 + \frac{\rho}{2}\|Z - (H + \Omega)\|_2^2 + \mathcal{I}_+(Z) \right\} \tag{6}$$

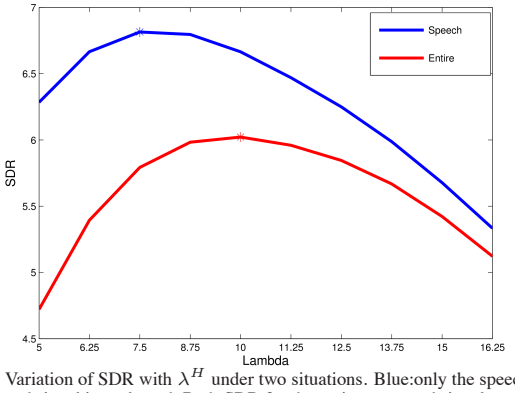

Fig. 1. Variation of SDR with λ^H under two situations. Blue:only the speech part of the separated signal is evaluated. Red: SDR for the entire separated signal.

Algorithm 1 Content-AdaptiveSaD-RPCA (CaSaD-RPCA)

Input: $Y, W, blocks$
Output: H, L
Initialization: $H = random; L = 0; Z = 0; \Omega = 0; t = 1$
while not converged **do**
  update H, Z:
  for each block l **do**
    $\lambda^H = \lambda_l^H$;
    $H_l^{t+1} = (W_l^\top W_l + \rho I_l)^{-1}(W_l^\top (Y_l - L_l^t) + \rho(Z_l^t - \Omega_l^t))$
    $Z_l^{t+1} = \mathcal{S}_{+\lambda^H/\rho}(H_l^{t+1} + \Omega_l^t)$
  end for
  $H^{t+1} = [H_1 H_2 \cdots H_{N_{\text{block}}}]$
  $Z^{t+1} = [Z_1 Z_2 \cdots Z_{N_{\text{block}}}]$
  update L:
  $U\Sigma V = svd(Y - WH^{t+1}); L^{t+1} = U\mathcal{S}_{\lambda^L}(\Sigma)V$
  update Ω:
  $\Omega^{t+1} = \Omega^t + H^{t+1} - Z^{t+1}$
  $t = t + 1$
end while

Without the term $\mathcal{I}_+(Z)$, the solution of (6) would be given by the soft-threshold operator [24, 25] [1]:

$$Z^{k+1} = \mathcal{S}_{\frac{\lambda^H}{\rho}}[H + \Omega]$$

The effect of the indicator function $\mathcal{I}_+(Z)$ is to project onto the first orthant, leading to:

$$Z^{k+1} = \max\left\{0, \mathcal{S}_{\frac{\lambda^H}{\rho}}[H + \Omega]\right\} \quad (7)$$

where the maximum is to be understood in the componentwise sense.

To incorporate side information in the decomposition, we consider a time segmentation of the magnitude spectrogram into N_{block} consecutive (non-overlapping) blocks of speech / non-speech (background noise) segments. We can represent the magnitude spectrogram as a concatenation of column-blocks $Y = [Y_1 Y_2 \cdots Y_{N_{\text{block}}}]$, the sparse layer as $Z = [Z_1 \cdots Z_{N_{\text{block}}}]$ and so on.

We can minimize the objective function with respect to each column-block separately. To guide the separation, we aim at setting a different value of $\lambda_l^H, l \in [1, N_{\text{block}}]$ for each block according to the speech activity side information. Note that we could further optimize the approach by choosing different ρ for each block, but in this work we hold ρ constant. For each block, the problem is equivalent to Eq. (6) and accordingly, the solution to the resulting problem:

$$Z_l^{k+1} = \min_{Z_l}\left\{\lambda_l^H \|Z_l\|_1 + \frac{\rho}{2}\|Z_l - (H_l + \Omega_l)\|_2^2 + \mathcal{I}_+(Z_l)\right\}$$

is given by:

$$Z_l^{k+1} = \max\left\{0, \mathcal{S}_{\frac{\lambda_l^H}{\rho}}[H_l + \Omega_l]\right\} \quad (8)$$

Using a large λ_l^H in blocks without speech will favor retaining all non-zero coefficients in the background layer. We detail here how we adapt the decomposition when exact speech activity location prior information is incorporated. Denoting by Ω_{speech} the set of time frames that contain speech, the values of λ_l^H are set as:

$$\forall\, l \in [1, N_{\text{block}}] \begin{cases} \lambda_l^H = \lambda_{\text{s}}^H \text{ if } Z_l \subset \Omega_{speech} \\ \lambda_l^H = \lambda_{\text{ns}}^H \text{ otherwise} \end{cases} \quad (9)$$

with $\lambda_{\text{ns}}^H > \lambda_{\text{s}}^H$ to enhance sparsity of Z when no speech activity is detected. In the evaluation, we will present experiments where the parameter λ^H is not binary, but more precisely designed to better suit the semantic information that is incorporated. The update rules of the CaSaD-RPCA algorithm are detailed in Algorithm 1. There, $\mathcal{S}_\lambda(\cdot)$ refers to the well-known soft-threshold operator [26][2], and $\mathcal{S}_{+\lambda}[\cdot]$ indicates the additional non-negative projection step after the soft-threshold step.

In Section 3, we investigate the results of content-adaptive SaD-RPCA using both exact and estimated speech activity side information. A noise-robust pitch tracking method based on subband autocorrelation classification (SAcC) is proposed in [27]. Pitch extraction and voicing detection is obtained by classifying the autocorrelations of a set of subbands from an auditory filterbank using an multilayer perceptron neural network. We use this algorithm to obtain speech activity information as it has been shown to be particularly effective for speech activity detection in high-noise situations [28].

3. EVALUATION

3.1. Parameters, Dataset, and Criteria

The proposed algorithm was evaluated with 400 noisy speech examples, totaling 3.5 hours of audio. The noisy signal were generated by adding clean speech to noise signals of different types and different signal-to-noise-ratios (SNRs). The clean speech was randomly collected from the BABEL Vietnamese language pack[3], a dataset of conversational telephone speech. These single conversation sides contain approximately equal proportions of active speech regions and silent gaps. The noise data were drawn from the AURORA dataset and other internet resources. We include 8 stationary noises – car, exhibition, restaurant, babble, train, subway, train, airport – and 3 transient noises – casino, keyboard, and birds. The test samples were mixed with noise at four SNRs from -10 to 5 dB. All signals were resampled to 8 kHz . The spectrograms were calculated using a window of 32 ms and a hop of 10 ms.

The speech dictionary W was learned from 10 conversation sides, each of about 10 minutes. Each side consisted of a different speaker, and these were disjoint from the speakers used to make the test samples. Sparse NMF with generalized KL-divergence [29] was used to generate the dictionary that contained 800 bases.

The non-adaptive SaD-RPCA model was used as the baseline, with $\lambda^L = 500$ and $\lambda^H = 10$, chosen empirically. Three different versions of the proposed CaSaD-RPCA algorithm were evalu-

[1] The scalar soft-thresholding (shrinkage) operator is defined as $\mathcal{S}_\epsilon[x]$:

$$\mathcal{S}_\epsilon[x] = \text{sgn}(x) \cdot \max(|x| - \epsilon, 0) = \begin{cases} x - \epsilon & \text{if} \quad x > \epsilon \\ x + \epsilon & \text{if} \quad x < -\epsilon \\ 0 & \text{otherwise} \end{cases}$$

where $x \in \mathbb{R}$ and $\epsilon > 0$. This operator can be extended to matrices by applying it element-wise.

[2] The scalar soft-thresholding (shrinkage) operator $\mathcal{S}_\lambda[x]$ is defined as : $\mathcal{S}_\lambda[x] = \text{sgn}(x) \cdot \max(|x| - \lambda, 0)$, where $x \in \mathbb{R}$ and $\lambda > 0$. It can be extended to matrices by applying it element-wise.

[3] IARPA Babel Program Vietnamese language collection release babel107b-v0.7, FullLP.

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	-12.79	-9.10	-7.16	-6.11	-6.67	-5.88
-5dB	-9.10	-3.44	-0.74	1.08	0.28	0.53
0dB	-2.95	3.16	4.25	5.16	4.35	4.83
5dB	2.04	6.81	7.11	7.57	7.46	8.03

Table 1. SDR values (in dB) for the whole utterance.

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	-12.79	-6.47	-3.74	1.29	-0.09	-0.64
-5dB	-7.91	-0.29	3.98	8.28	9.59	7.83
0dB	-2.95	6.24	12.54	15.38	16.15	14.89
5dB	2.04	10.24	20.11	21.69	21.37	20.14

Table 2. SIR for the whole utterance

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	43.79	2.22	2.03	0.31	-0.44	0.30
-5dB	43.91	3.11	4.02	3.48	2.42	3.20
0dB	44.00	7.31	6.14	6.19	5.37	6.06
5dB	44.19	10.22	7.99	7.96	7.95	8.68

Table 3. SAR for the whole utterance, for all systems at various SNRs, averaged across all noise types. Ori is the SDR value of original noisy speech. SaD is the baseline system, CaSaD_GT is the adaptive version using ground truth speech activity information, CaSaD_b uses binary estimated speech activity, and CaSaD_m uses multi-level estimated speech activity. MMSE is the comparison algorithm [30].

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	-9.17	-5.82	-4.25	-3.92	-4.19	-4.00
-5dB	-4.30	-0.82	1.41	1.69	1.12	1.48
0dB	0.66	4.53	5.38	5.69	4.75	5.25
5dB	5.65	8.00	7.61	8.09	7.71	8.29

Table 4. SDR for the speech part only.

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	-9.17	-2.74	-0.70	-0.34	1.39	0.80
-5dB	-4.30	-3.66	6.29	6.35	9.40	7.65
0dB	0.66	8.21	13.42	12.96	14.51	13.15
5dB	5.65	11.62	19.34	18.77	18.82	17.48

Table 5. SIR values (in dB) for the speech part

	Ori	MMSE	SaD	CaSaD_GT	CaSaD_b	CaSaD_m
-10dB	287.13	2.43	2.91	2.91	0.80	1.65
-5dB	286.85	3.16	5.13	5.35	3.29	4.42
0dB	287.56	7.97	6.91	7.35	5.99	6.96
5dB	289.16	11.12	8.17	8.76	8.39	9.34

Table 6. SAR values (in dB) for the the speech part only.

ated. In each, the value of value λ^L remained the same while the value of λ^H was adapted through the decomposition according to speech activity information. First was CaSaD-RPCA with the exact speech activity, using manually annotated ground-truth (CaSaD_GT) and $\lambda_l^H = \lambda^H$ for speech regions and $\lambda_l^H = 2\lambda^H$ for noise only (silent) regions. For the two remaining conditions, estimated speech activity from SAcC was used. SAcC outputs the posterior probability of voicing for each frame in the noisy speech; this probability was median-filtered over a 50 frame (500 ms) window, and the smoothed estimate was used two ways: First, CaSaD_b thresholds the voice activity probability into a binary value for each frame, and adapts the λ_l^H parameter as in the case of CaSaD_GT. The final configuration (CaSaD_m) uses instead a ten-level quantization of the speech activity probability, and correspondingly sets λ_l^H to ten equally-spaced values between λ^H and $2\lambda^H$. We also include comparison with the classic LogMMSE estimation [30] (MMSE).

The widely used Signal-to-Distortion Ratio (SDR) from the BSS_EVAL package [31] was used as the evaluation criteria; a larger score indicates better performance. To accurately evaluate the performance of the algorithms, the SDR of both the entire utterance and the speech-only parts are reported. The result are shown in Tables 1 and 4. Paired sample t-tests at the 5% significance level were performed to determine the statistical significance of the results.

The algorithms are implemented in MATLAB and performed on a MacBook Pro Intel Core 2 at 2.4GHz with 2GB RAM. 30s are needed to process 40s of noisy speech for both SaD and CaSaD.

3.2. Results and Discussion

Global enhancement results: As we can see from Table 1 all the three comparison algorithms outperformed the baseline models on the entire test and in all SNRs conditions. Statistical tests show that the difference in the results is significant. In all experiments, for a given constant value λ_s in Eq. (9), setting $\lambda_{ns}^H > \lambda_s^H$ always improves the results. This shows that a structurally-informed adaptive regularization parameter allows improved speech enhancement. However, note that artifacts may be introduced in low SNR conditions, as can be seen in the SAR results, Tab. 3.1.

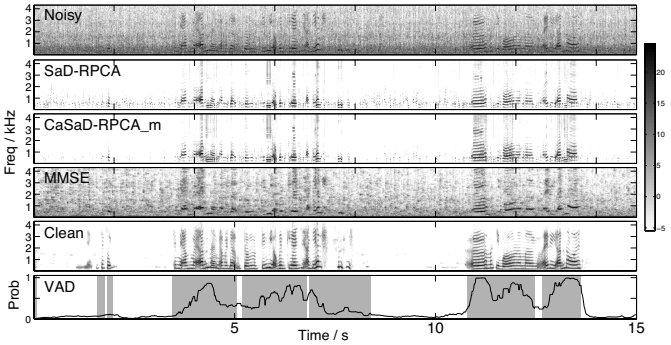

Fig. 2. Example decomposition. The top pane shows the spectrogram of speech mixed with babble noise at 0 dB. The separated speech part of the baseline model (SaD-RPCA) is shown in the second panel, and the third pane shows the sparse component of the proposed CaSaD-RPCA_m model (CaSaD-RPCA_GT and CaSaD-RPCA_b are similar). For comparison, the fourth pane shows log-MMSE enhancement. The clean speech appears in the fifth pane, with the smoothed estimated voice activity shown in the bottom pane. Shaded regions correspond to true speech activity (derived from the clean signal).

Speech parts only enhancement results: We expect the noise-only (silent) parts of the utterance to be improved with the CaSad-RPCA algorithm, since the side information directly indicates these regions where the foreground (sparse) components should be avoided; this can be clearly seen in Fig. 2. However, the improvements under the proposed model are not limited to speech-inactive sections. Tab. 4 shows that by using the adaptive algorithm, the speech-active segments are also better enhanced. Indeed, apart from the binary model at -5dB and 0dB, the separation is uniformly significantly better when measured on the speech parts alone. This indicates that side information helps not only to determine the silent gaps, but also enables improved recovery of the speech, presumably because the low-rank noise model is a better match to the actual noise.

Ground truth versus estimated speech activity location: The results show that imperfect, estimated speech activity information still allows an improvement, although not as much as with ground-truth speech activity information. The decrement in performance is mainly due to non-speech segments being classified as speech segments. Results obtained with CaSaD-RPCA_m suggest that, when using estimated speech activity as side information, a multiple-level adaptive parameter helps reduce the impact of misclassified frames.

4. CONCLUSION

In this work, we have explored a novel framework for speech enhancement based on a combination of Robust PCA and learned target source models. Our approach incorporates side information to adapt the decomposition to the local content of the audio excerpt. Our experiments show that the proposed model is superior to existing approaches when applied to entire conversation sides that may exhibit large variations in the speech content. We continue to investigate mechanisms to improve the quality of the separated target speech, for instance by incorporating other information, such as knowledge of the speaker's gender which could help guide the sparse layer towards appropriate speech model bases.

5. REFERENCES

[1] S. Brutzer, B. Hoferlin, and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *CCVPR*, 2011, pp. 1937–1944.

[2] E.J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, Article 11, 2011.

[3] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky, "Sparse and low-rank matrix decompositions," in *Sysid*, 2009.

[4] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *ICCV*, 2011, pp. 2439–2446.

[5] Z. Zeng, T.H. Chan, K. Jia, and D. Xu, "Finding correspondence from multiple images via sparse and low-rank decomposition," in *ECCV*, 2012, pp. 325–339.

[6] F. Yang, H. Jiang, Z. Shen, W. Deng, and D.N. Metaxas, "Adaptive low rank and sparse decomposition of video using compressive sensing," *CoRR*, vol. abs/1302.1610, 2013.

[7] Y. Peng, A. Ganesh, J. Wright, and Y. Xu, W. andMa, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, 2012.

[8] P.S. Huang, S.D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing voice separation from monaural recordings using robust principal component analysis," in *ICASSP*, 2012.

[9] Y.H. Yang, "On sparse and low-rank matrix decomposition for singing voice separation," in *MM*, 2012, pp. 757–760.

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

[11] J. Huang, X. Zhang, Y. Zhang, X. Zou, and L. Zeng, "Speech denoising via low-rank and sparse matrix decomposition," *ETRI Journal*, vol. 36, no. 1, pp. 167–170, 2014.

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

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

[14] D.L. Sun and G.J. Mysore, "Universal speech models for speaker independent single channel source separation," in *ICASSP*, 2013.

[15] M. Fazel, *Matrix Rank Minimization with Applications*, Ph.D. thesis, Dept of Elec. Eng., Stanford Univ., 2002.

[16] B. Recht, M. Fazel, and P.A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.

[17] E.J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[18] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," Tech. Rep. UILU-ENG-09-2214, UIUC Tech. Rep., 2009.

[19] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," Tech. Rep. UILU-ENG-09-2215, UIUC, 2009.

[20] Xiaoming Yuan and Junfeng Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *Preprint*, pp. 1–11, 2009.

[21] H. Zou, "The adaptive lasso and its oracle properties," *J. Am. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[22] D. Angelosante and G. Giannakis, "Rls-weighted lasso for adaptive estimation of sparse signals," in *ICASSP*, 2009, pp. 3245–3248.

[23] Y. Grandvalet, "Least absolute shrinkage is equivalent to quadratic penalization," in *ICANN 98*, L. Niklasson, M. Boden, and T. Ziemke, Eds., Perspectives in Neural Computing, pp. 201–206. Springer London, 1998.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[25] S. Chen, L. David, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[26] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[27] B.S. Lee and D.P.W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *INTERSPEECH*, 2012.

[28] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. Hansen, A. Janin, B.-S. Lee, Y. Lei, V. Mitra, N. Morgan, S.O. Sadjadi, T.J. Tsai, N. Scheffer, L. N. Tan, and B. Williams, "All for one: Feature combination for highly channel-degraded speech activity detection," in *Proceedings of Interspeech*, 2013.

[29] M.N. Schmidt and R.K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proceedings of Interspeech*, 2006, pp. 2614–2617.

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

[31] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.