Unsupervised Deep Learning of Sparse Signals against Low-Rank Backgrounds with Application to Online Lung Sound Separation

Takumi Onomichi*, Tomoya Sakai, and Yasushi Obase

Graduate School of Engineering, Nagasaki University, Japan;
Email: tsakai@cis.nagasaki-u.ac.jp (T.S.), obaseya@nagasaki-u.ac.jp (Y.O.)
*Correspondence: b317008@cis.nagasaki-u.ac.jp (T.O.)

***Abstract*—This paper presents an unsupervised deep learning approach to online sparse signal extraction. For a batch of input vectors of low-rank and sparse components, a U-Net-based model is trained with a combination of nuclear and ℓ_1 norms as a loss function so as to encode and decode only the sparse component of each input vector. Since the model learns general structures common to the sparse components of the input vectors, it has the potential to distinguish them not only from low-rank but also from any background. After the training, the model can extract a learned sparse component from any input signal with much lower computational complexity than iterative algorithms of robust principal component analysis (RPCA). In an application to respiratory auscultation, continuous adventitious sounds can be extracted as the sparse components of the spectrograms of input auscultation signals. It is experimentally demonstrated that a well-generalized model that outperforms RPCA can be obtained by learning only a few hundreds of windowed signals of an auscultatory sound.**

***Keywords*—nuclear loss, dual frame U-Net, low-rank and sparse model, lung sound analysis**

I. Introduction

We aim to detect informative signal features of sequential data. While recurring and redundant features of temporal signals are likely derived from structural events, sparse outlying features indicate unusual/abnormal events of interest. Sequential data with redundant and sparse features can be represented by a low-rank and sparse (L+S) model in many cases.

In fact, the L+S model applies well to a variety of time series data such as surveillance video (background+ foreground) [1, 2], optical-flow sequences (egomotion+ object motion) [3], music (accompaniment+singing voice) [4], etc. Respiratory auscultation sounds are also a typical example in medical practice [5–11]. Repeating breath sounds have low-rank structures in a time-frequency domain. In contrast, continuous adventitious sounds, e.g., wheezes and rhonchi, consist only of a few harmonic partials with fluctuating fundamental frequencies, and their spectrograms show sparse and curved patterns.

The robust principal component analysis (RPCA) [1, 12] is a successful technique for capturing the low-rank structure with its principal components while excluding sparse outliers. One of the limitations of RPCA is that it learns only principal components, which is particularly helpless in many applications where the sparse component is of interest. For example, the low-rank background and sparse foreground obtained by RPCA from a video are useless in the background subtraction of any other video from a different viewpoint even if the foreground is always common objects such as pedestrians. In lung sound analysis, we would like to have a trainable model that learns from auscultatory sound signals the features of structured sparse spectrogram components of wheezes and rhonchi in an unsupervised manner, just as a doctor or nurse would from the auscultation experience.

In this paper, we propose unsupervised deep learning of sparse signal components that exploits low-rank and sparse priors for training. We employ a popular hourglass-like model with convolutional layers because it is reasonable for encoding and decoding sparse components. We train our model using the sum of nuclear and ℓ_1 norms, i.e., the objective function of RPCA as the loss function to minimize with respect to the model parameters. Among the recent methods of deep learning-based RPCA [13–16], ours is the most concise and practical way to separate sparse signals from mixtures with low-rank backgrounds. Our model and its unsupervised training are presented in section II.

Once our model has learned enough sparse components, it can detect and extract them from any mixtures, not just those with low-rank ones. Therefore, we do not have to retrain the model to adapt to new backgrounds. This enables computationally efficient online separation of target signals whose features have been learned as the sparse components of training data sequences. In section III, we experimentally show that our model can be generalized well by learning only a few hundreds of windowed training signals in the context of continuous adventitious sound separation.

Manuscript received July 12, 2022; revised September 7, 2022; accepted October 16, 2022.

doi: 10.18178/ijsps.11.1.1-6

II. Unsupervised Deep Learning for Online Separation of Sparse Signals

A. Low-Rank and Sparse (L+S) Sequential Data Model

Let $\{\boldsymbol{d}^{(j)}\}$ $(j = 1, \ldots, n)$ be a sequence of m-dimensional vector data. We assume that $\{\boldsymbol{d}^{(j)}\}$ is a sum of two sequences $\{\boldsymbol{l}^{(j)}\}$ and $\{\boldsymbol{s}^{(j)}\}$ with redundant (i.e., linearly dependent) and sparse features, respectively. The linear dependence of $\{\boldsymbol{l}^{(j)}\}$ can be quantified as the low-rankness of a m by n matrix $\boldsymbol{L} = [\boldsymbol{l}^{(1)}, \ldots, \boldsymbol{l}^{(n)}]$. The sparsity of $\{\boldsymbol{s}^{(j)}\}$ can be measured as the number of nonzero entries of $\boldsymbol{S} = [\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(n)}]$.

For a given matrix $\boldsymbol{D} = [\boldsymbol{d}^{(1)}, \ldots, \boldsymbol{d}^{(n)}]$, its L+S model is the sum of the low-rank matrix $\boldsymbol{L}$ and the sparse matrix $\boldsymbol{S}$. Fitting the L+S model to $\boldsymbol{D}$ can be posed as the following convex optimization problem [1].

$$\underset{(\boldsymbol{L},\boldsymbol{S})}{\text{Minimize}} \ \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1 \ \text{ subject to } \ \boldsymbol{D} = \boldsymbol{L} + \boldsymbol{S} \quad (1)$$

Here, the nuclear norm of $\boldsymbol{L}$, $\|\boldsymbol{L}\|_*$, is defined as the sum of the singular values of $\boldsymbol{L}$. The matrix ℓ_1 norm of $\boldsymbol{S}$, $\|\boldsymbol{S}\|_1$, is defined as the sum of the absolute values of the matrix entries. Minimizing the nuclear norm $\|\boldsymbol{L}\|_*$ and ℓ_1 norm $\|\boldsymbol{S}\|_1$ promotes the low-rankness and sparseness of $\boldsymbol{L}$ and $\boldsymbol{S}$, respectively. The hyperparameter $\lambda > 0$ balances these norms in the minimization.

B. Dual Frame U-Net Model for Sparse Feature Learning

We design a deep neural network model that sequentially separates learned sparse components. The model architecture is shown in Fig. 1. Our model takes a column vector $\boldsymbol{d}^{(j)}$ of $\boldsymbol{D}$ as input, and outputs the corresponding low-rank component $\boldsymbol{l}^{(j)}$ and sparse component $\boldsymbol{s}^{(j)}$. The encoding and decoding to produce the sparse component $\boldsymbol{s}^{(j)}$is handled by a dual frame U-Net [17] with a set Θ of learnable parameters. The output of the low-rank component is simply computed as $\boldsymbol{l}^{(j)} = \boldsymbol{d}^{(j)} - \boldsymbol{s}^{(j)}$.

The dual frame U-Net has an hourglass structure with convolutional layers. A unique aspect of dual frame U-Net is that the features right after pooling in the encoder are subtracted from the features right before upsampling in the decoder at each level. While the fine structure of the input $\boldsymbol{d}^{(j)}$ is lost at each pooling of the encoder, the remaining coarse structure is processed for feature extraction down to the deeper levels. The dual frame U-Net is an improved model of U-Net [18] that avoids this imbalance of fine and coarse structures, and is therefore suitable for learning sparse components with fine structure consisting of a small number of non-zero entries.

Note that the dual frame U-Net in Fig. 1 consists of 1D convolution layers for lung sound analysis, where $\{\boldsymbol{d}^{(j)}\}$ is a sequence of the Fourier transforms of windowed signals and $\boldsymbol{D}$ is a spectrogram, i.e., the short time Fourier transform (STFT) of auscultatory sound. For analysing image sequences, one can design the same neural network architecture using 2D convolution layers, where $\boldsymbol{d}^{(j)}$ represents a vector of pixel values of the j-th frame. The proper number of channels and layers depend on an application.

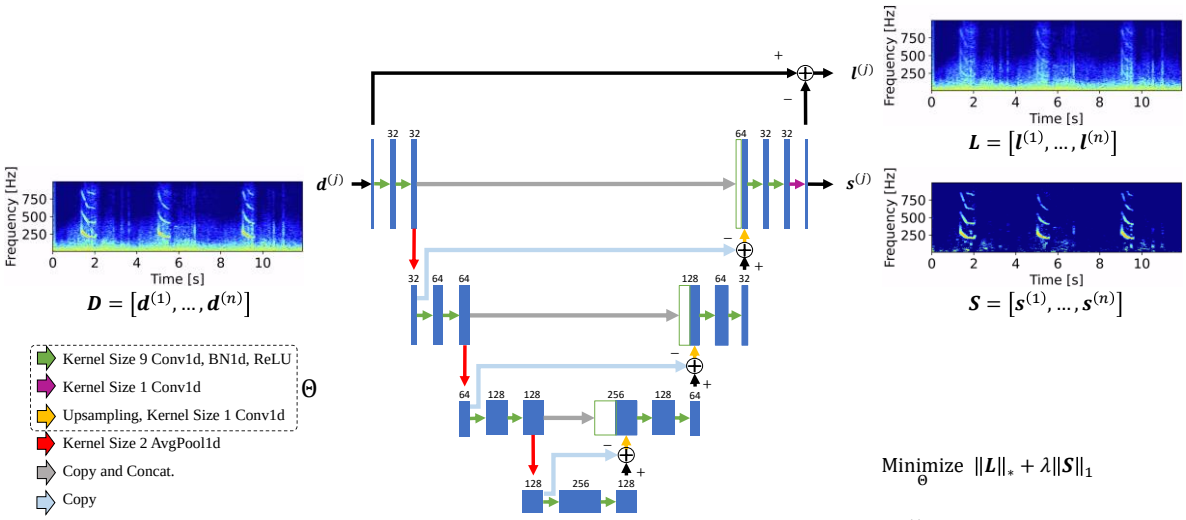

Figure 1. Our U-Net-based model (for online lung sound separation). Given an $\boldsymbol{m}$-dimentional column vecter $\boldsymbol{d}^{(j)}$ $(\boldsymbol{j} = \boldsymbol{1}, \ldots, \boldsymbol{n})$ of a $\boldsymbol{m}$ by $\boldsymbol{n}$ matrix $\boldsymbol{D}$ (STFT of an auscultatory sound), the dual frame U-Net [17] outputs its sparse component $\boldsymbol{s}^{(j)}$ of $\boldsymbol{S}$ (STFT of wheezes/rhonchi only). The model is trained so that the matrices $\boldsymbol{L}$ and $\boldsymbol{S}$ storing the output $\{\boldsymbol{l}^{(j)}\}$ and $\{\boldsymbol{s}^{(j)}\}$ are as low-rank and sparse as possible by minimizing the nuclear and $\boldsymbol{\ell_1}$ loss, respectively.

C. Unsupervised Training with Nuclear and ℓ_1 Norms

To train our model in Fig. 1, the same objective function in Eq. (1) is used as the loss function. For a series of training inputs $\{\boldsymbol{d}^{(j)}\}$ $(j = 1, \ldots, n)$, we construct matrices $\boldsymbol{L}$ and $\boldsymbol{S}$ with the corresponding batches of model outputs, $\{\boldsymbol{l}^{(j)}\}$ and $\{\boldsymbol{s}^{(j)}\}$, respectively as their columns. Note that $\boldsymbol{D} = \boldsymbol{L} + \boldsymbol{S}$ holds due to the model architecture. We optimize the parameter Θ so that the loss function is minimized:

$$\underset{\Theta}{\text{Minimize}} \ \|\boldsymbol{L}\|_* + \lambda\|\boldsymbol{S}\|_1. \quad (2)$$

This is unsupervised learning; the output data that the model should produce for the input do not need to be

explicitly given. It is theoretically reasonable to choose λ to be $\lambda_0 = 1/\sqrt{\max(m,n)}$ [1]. We recommend a larger value of λ to avoid learning irrelevant features other than those of the sparse components of interest.

For convenience, we use PyTorch [19], an autodiff framework with the singular value decomposition (SVD) and the optimizer Adam [20] available, to implement our L+S deep learning model and the loss function in Eq. (2).

III. Lung Sound Separation: Experimental Comparison with RPCA

We experimentally show that our model in Fig. 1 can be trained to detect and output continuous adventitious sounds (wheezes/rhonchi). The left column of Fig. 2 displays the waveforms and spectrograms of respiratory auscultation sounds used for the input to our model. The duration of every signal is about 12 seconds with a sampling frequency of 11.025 kHz. An input sequence $\{\boldsymbol{d}^{(j)}\}$ is given as positive frequency components of the Fourier transforms of the windowed input signals. The window size and hop length are 2048 and 512, respectively. Consequently, $\boldsymbol{D}$ is a complex matrix of size 1024×257. Note that the spectrograms in Fig. 2 are displayed between 0 and 1kHz in dB.

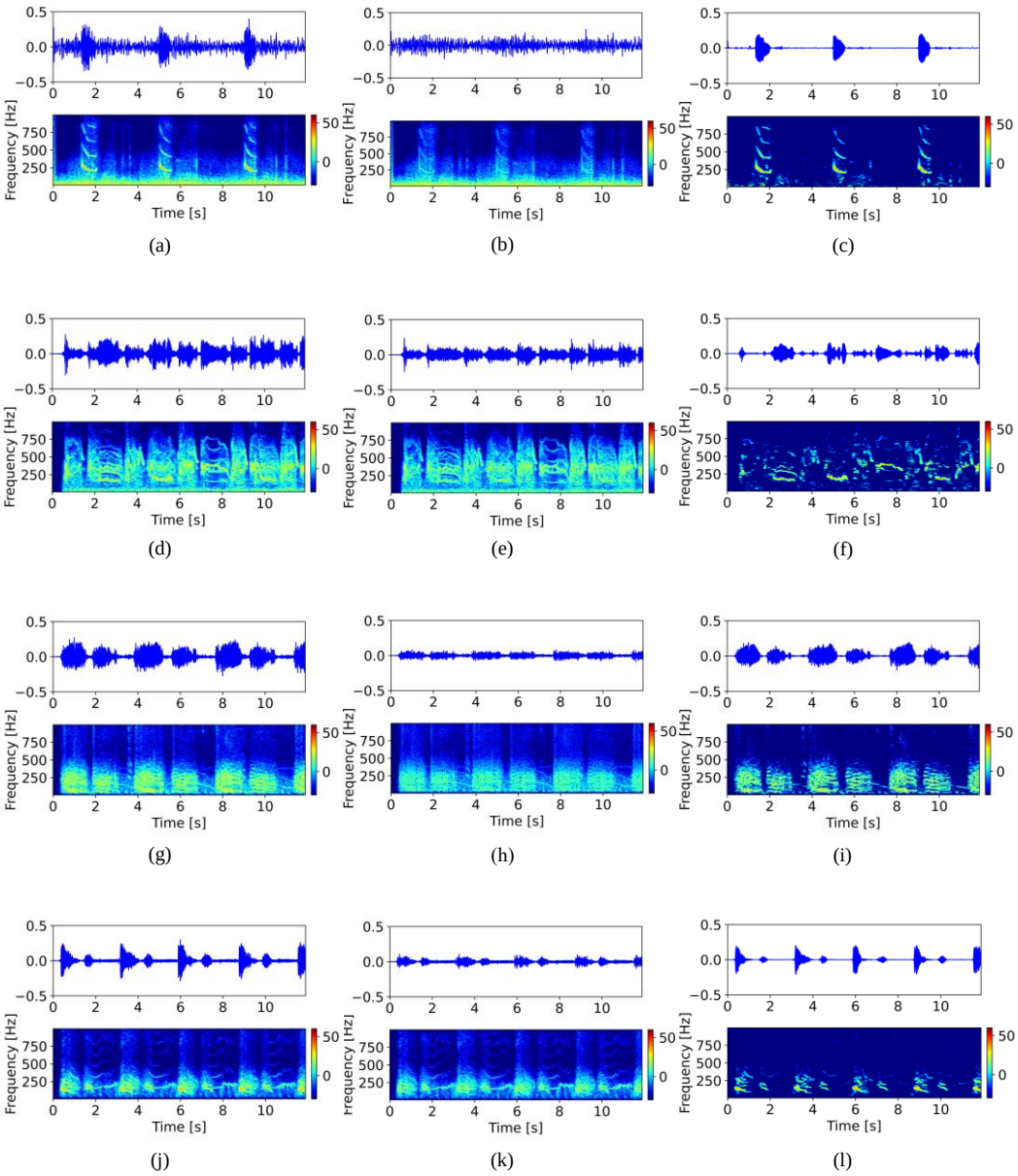

Figure 2. Lung sound separation by our unsupervised deep learning method. The left column shows waveforms and their spectrograms of adventitious respiratory sounds (wheezes and/or rhonchi): (a) training sound T1, (d) training sound T2, (g) test sound Q1, and (j) test sound Q2. Each spectrogram is used as $\boldsymbol{D}$ for the input to our model. The middle and right columns show the outputs, respectively $\boldsymbol{L}$ and $\boldsymbol{S}$, of our model. The waveforms of the model outputs are obtained by the inverse STFT of the spectrograms $\boldsymbol{L}$ and $\boldsymbol{S}$ shown below each.

We trained our model on input sounds T1 and T2 respectively shown in Fig. 2(a) and (d) using Adam with learning rate 10^{-3} for 800 epochs. It took about ten minutes on a single NVIDIA Tesla K80 GPU. We set $\lambda =$

$3\lambda_0 \approx 0.094$ because we found that a smaller value would have resulted in noticeable false positive extraction for training data T1 and T2, and the model would learn features that are not related to the continuous adventitious sounds.

A. *Result of Our Method*

The middle and right columns of Fig. 2 display the waveforms and spectrograms of the low-rank and sparse output signals, respectively. We have confirmed that the model after training can output the low-rank and sparse spectrograms, $\boldsymbol{L}$ and $\boldsymbol{S}$, as shown in Fig. 2(b) and (c) for T1, (e) and (f) for T2. This model was able to learn the adventitious sounds enough to output (c) and (f) for the training input T1 and T2, respectively. Although the ground-truth waveforms and spectrograms of breath and continuous adventitious sounds are not available by auscultation, rhonchi contained in the training input T1 are clearly represented as the fragments of curves in the sparse spectrogram of (c). There remain weak curves in the low-rank spectrogram of (e), but the model almost never mis-extracts non-curve components as sparse components.

We tested our model on input sounds Q1 and Q2 respectively shown in Fig. 2(g) and (j). Both inputs contain rhonchi and their main frequency components are well extracted as the sparse components shown in (i) for Q1 and (l) for Q2. The extracted sparse components tend to form continuous curves in the spectrograms. Note that the forward propagation computation time required for a test output of 12 seconds was less than two hundred milliseconds on the GPU and several seconds on a single core of a modern CPU. This is enough for real-time processing.

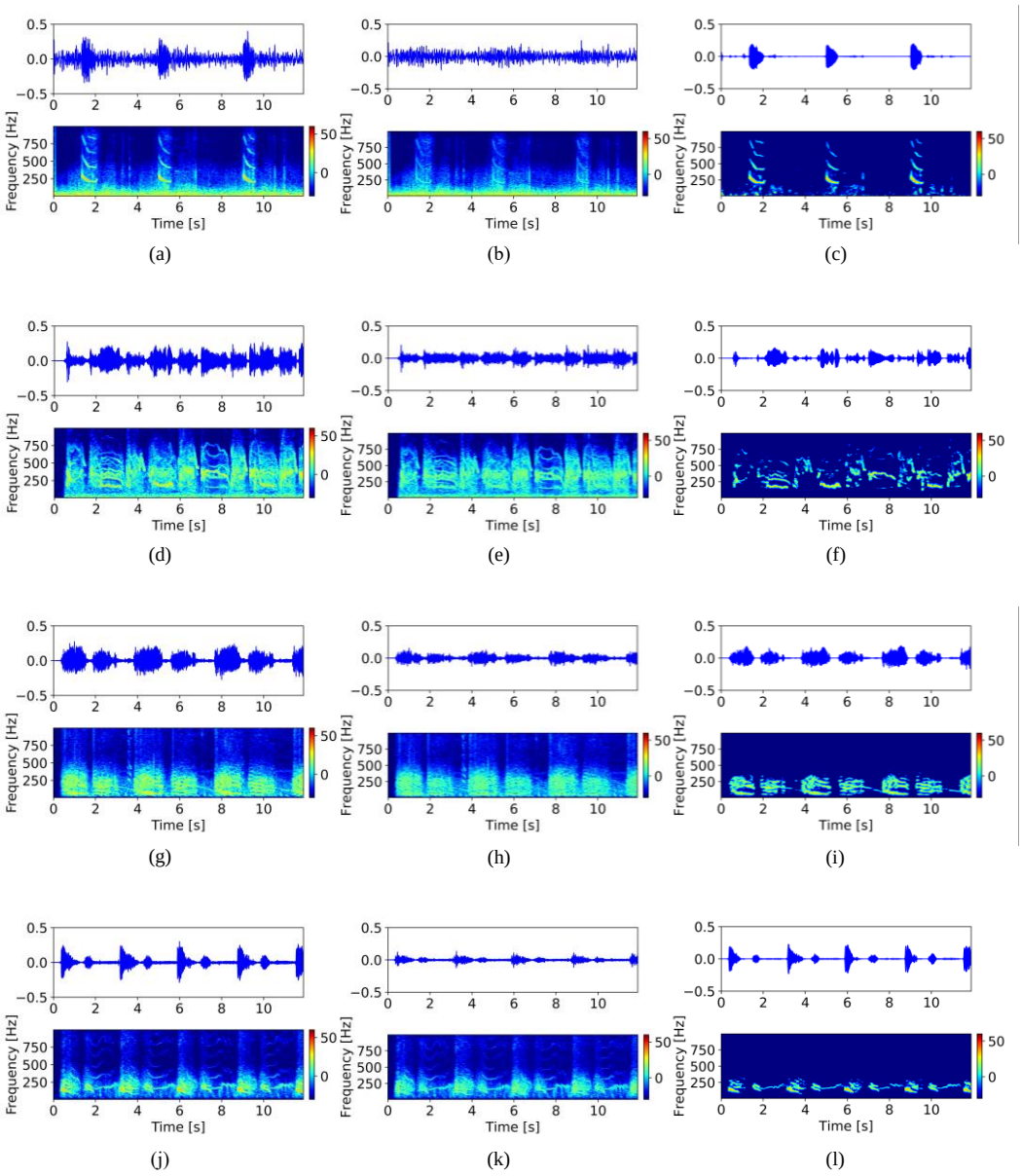

Figure 3. Lung sound separation by robust principal component analysis (RPCA). The left column shows waveforms and their spectrograms as in Fig. 2. Each spectrogram is used as $\boldsymbol{D}$ in Eq. (1). The waveforms in the middle and right columns are obtained from low-rank and sparse spectrograms, $\boldsymbol{L}$ and $\boldsymbol{S}$, shown below each, respectively.

B. Comparsion with RPCA Ressults

Figure 3 shows the results for the same auscultation sounds by RPCA. The left columns of Fig. 3 display the waveforms and spectrograms of the input signals T1, T2, Q1, and Q2 in the same way as Fig. 2. The waveforms and spectrograms of the corresponding low-rank and sparse components obtained by RPCA are shown in the middle and right columns. We had to roughly find and set an adequate value of the hyperparameter λ in Eq. (1) for each sound: $\lambda = 3\lambda_0$ for T1 and T2, and $\lambda = 2\lambda_0$ for Q1 and Q2. When λ was set larger than these values, only a few parts of wheezes/rhonchi were extracted; when λ was smaller, the spectrogram $\boldsymbol{S}$ lost sparsity and $\boldsymbol{S} \approx \boldsymbol{D}$. One advantage of using our model over RPCA is that it avoids this hyperparameter tuning during inference.

For T1 and T2, the results (b), (c), (e), and (f) of RPCA in Fig. 3 are very similar to those of our method in Fig. 2. Figure 3(f) has more isolated instantaneous frequency components that are not relevant to the wheezes/rhonchi with curved patterns. While RPCA individually selects the instantaneous frequency components of the sparse spectrogram, our model decodes the sparse components by the convolution layers. This implies that our model is better at capturing structured sparse components, making the main difference between Figs. 2(i) and 3(i) and Figs. 2(j) and 3(j).

RPCA finds more curved patterns along with false positive extraction of irrelevant components for Q1 and Q2. RPCA could not, however, reduce this false positive extraction without increasing false negative extraction. It is also hard to obtain the result of RPCA in real-time. An ADMM-based algorithm took more than twenty times longer to converge than inference by our model. Although our model misses some weak harmonic partials of rhonchi for Q1 and Q2, it can efficiently find in input sounds the structured sparse components learned from T1 and T2. It is expected that increasing training on auscultation sounds with low-rank and sparse structure will reduce false negative extraction.

IV. Conclusions

Our L+S neural network modeling enables unsupervised deep learning of sparse components in the mixtures with low-rank components. After training only on a small number of data, the model gets generalized and is capable of fast online separation of signals with similar features as the learned sparse components. We discussed its advantages over RPCA in the application to lung sound separation. Even though quantitative evaluation is impossible in this application because ground truth is unavailable, we have confirmed that it is possible to build a deep neural network for signal processing that are orders of magnitude faster than RPCA.

Further research should address quantitative evaluation on the generalization power to verify the advantages of our model having learned features. It is highly beneficial to apply not only the ℓ_1 norm but also other sparsity-inducing norms such as the nuclear norm to deep learning. The use of these norms as loss functions could realize unsupervised deep learning to yield highly generalized models even with a small amount of training data.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Tomoya Sakai and Yasushi Obase conducted the research; Takumi Onomichi analyzed the data; Takumi Onomichi and Tomoya Sakai wrote the paper; all authors had approved the final version.

Funding

This research was partially supported by TERUMO LIFE SCIENCE FOUNDATION (2021, Obase Y and Sakai T) and JSPS KAKENHI Grant (19H04177, Sakai T).

References

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11:1-11:37, 2011.

[2] C. Guyon, T. Bouwmans, E.-h. Zahzah, "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis," in *Principal Component Analysis*, P. Sanguansat, Ed., 2012, pp. 223-238.

[3] T. Sakai and H. Kuhara, "Separating background and foreground optical flow fields by low-rank and sparse regularization," in *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 1523-1527.

[4] 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. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 57-60.

[5] R. Mikami, M. Murao, D. W. Cugell, J. Chrétien, P. Cole, J. Meier-Sydow, R. L. H. Murphy, and R. G. Loudon, *International symposium on Lung Sounds. Synopsis of proceedings*, vol. 92, no. 2, pp. 342-345, 1987.

[6] H. Pasterkamp, C. Carson, D. Daien, and Y. Oh, "Digital respirosonography: New images of lung sounds," *Chest*, vol. 96, pp. 1405-1412, 1989.

[7] H. Pasika and D. Pengelly, "Lung sound crackle analysis using generalised time-frequency representations," *Medical and Biological Engineering and Computing*, vol. 32, pp. 688-690, 1994.

[8] T. Kaisia, A. Sovijärvi, P. Piirilä, H. Rajala, S. Haltsonen, and T. Rosqvist, "Validated method for automatic detection of lung sound crackles," *Medical and Biological Engineering and Computing*, vol. 29, no. 5, pp. 517-521, 1991.

[9] S. A. Taplidou and L. J. Hadjileontiadis, "Wheeze detection based on time-frequency analysis of breath sounds," *Computers in Biology and Medicine*, vol. 37, pp. 1073-1083, 2007.

[10] T. Sakai, H. Satomoto, S. Kiyasu, and S. Miyahara, "Sparse representation-based extraction of pulmonary sound components from low-quality auscultation signals," in *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 509-512.

[11] T. Sakai, S. Miyahara, and S. Kiyasu, "Unmixing three types of lung sounds by convex optimization," in *Proc. 2016 23rd International Conference on Pattern Recognition*, 2016, pp. 2884-2888.

[12] T. Bouwmans and E. H. Zahzah, "Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance," *Computer Vision and Image Understanding*, vol. 122, pp. 22-34, 2014.

[13] F. Bahri, M. Shakeri, and N. Ray, "Online illumination invariant moving object detection by generative neural network," in *Proc.*

11th Indian Conference on Computer Vision, Graphics and Image Processing, 2018, pp. 1-8.
[14] O. Solomon, R. Cohen, Y. Zhang, Y. Yang, Q. He, J. Luo, R. J. G. van Sloun, and Y. C. Eldar, "Deep unfolded robust PCA with application to clutter suppression in ultrasound," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1051-1067, 2020.
[15] C. Herrera, F. Krach, A. Kratsios, P. Ruyssen, and J. Teichmann, "Denise: Deep learning based robust PCA for positive semidefinite matrices," *arXiv:2004.13612*, 2020.
[16] B. Rezaei, A. Farnoosh, and S. Ostadabbas, "G-LBM: Generative low-dimensional background model estimation from video sequences," in *Proc. European Conference on Computer Vision*, Springer, 2020, pp. 293-310.
[17] Y. Han and J. C. Ye, "Framing U-Net via deep convolutional framelets: Application to sparse-view CT," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1418-1429, 2018.
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234-241.
[19] A. Paszke *et al*., "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8026-8037.
[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd International Conference on Learning Representations*, 2015.

Takumi Onomichi received his B. Eng. in 2021 from the Computer and Information Science Program, School of Engineering, Nagasaki University, Japan. He is currently a graduate student pursuing a master of engineering under the supervision of Assoc. Prof. Dr. Tomoya Sakai at Graduate School of Engineering, Nagasaki University. His research focus includes applications of deep learning to biosignal processing.

Tomoya Sakai is an associate professor at the Graduate School of Engineering and School of Information and Data Sciences, Nagasaki University, Japan. He received his M. Eng. And Dr. Eng. from Chiba University in 1998 and 2001, respectively. He was an assistant professor at IMIT, Chiba University from 2001 to 2010. He is a member of IEEE, IPSJ and IEICE. He served from 2014 until 2017 as an Associate Editor of the IEICE Transactions on Fundamentals of Electronics. His research interests range from signal processing to machine learning for computer vision and pattern recognition with a particular focus on sparsity-aware approaches to high-dimensional data analysis.

Yasushi Obase is an associate professor at Department of Respiratory Medicine, Nagasaki University Graduate School of Biomedical Sciences. He received his PhD from Nagasaki University in 2001. He was a postdoctoral researcher at Skin and Allergy Hospital, Helsinki University from 2001 to 2004. He was Assistant Professor at Kawasaki Medical School from 2004 to 2013. His research interests range from respiratory disease diagnosis by lung sounds to treatment with biochemic medicines.