DCASE 2020 TASK2: ANOMALOUS SOUND DETECTION USING RELEVANT SPECTRAL FEATURE AND FOCUSING TECHNIQUES IN THE UNSUPERVISED LEARNING SCENARIO

Jihwan Park and Sooyeon Yoo

Advanced Robot Research Laboratory, LG Electronics, Seoul, South Korea
{jihwan.park, sooyeon.yoo}@lge.com

ABSTRACT

In this paper, we propose an improved version of the anomalous sound detection (ASD) system for noisy and reverberant conditions, which was submitted to DCASE 2020 Challenge Task2. The improved system consists of three phases: feature extraction, autoencoder (AE) model, and focusing techniques. In the feature extraction phase, we used spectrograms instead of log-mel energies for more effective distinction of normal and abnormal machine sounds, and validated this feature for the baseline autoencoder model and interpolation DNN (IDNN). We also applied the focusing techniques in both train and evaluation phases, which focuses on machine-adaptive ranges of reconstructed errors for performance improvements. Through experiments, we found that our proposed ASD system outperforms baseline methods under the unsupervised learning scenario. The performance improvement was especially remarkable for non-stationary sounds; above 95% of AUC score was achieved for slider and valve sounds with the proposed system.

Index Terms— DCASE Challenge 2020 Task2, Anomalous sound detection, Unsupervised learning, Deep neural networks, Autoencoder, Focused back-propagation, Focused hypothesis test

1. INTRODUCTION

Recently, researches on unsupervised learning-based anomalous sound detection (ASD) is getting attention. For the training and fair evaluation of sound detection algorithms, a large-scale database is essential. However, due to the rarity and diversity of real-world anomalous sounds, it is difficult to create or collect massive patterns of such sounds. In consequence, we have to detect unknown anomalous machine sounds which have patterns not reflected in given training data.

Most ASD systems adopt outlier detection techniques because it is difficult to collect sufficient amount of anomalous machine sound data. In [1], deep neural network-based autoencoders (AE) were used to build up an ASD system. In their system, acoustic features were extracted from the encoder part, and the input vector was reconstructed at the decoder part of the AE. Using the reconstruction error defined as the mean squared error (MSE) between the input and reconstructed vectors, statistical hypothesis test was conducted with a predefined threshold value [2]. In [3]-[4], AE-based an acoustic feature extractor was optimized to maximize the true positive rate under an arbitrary false positive rate, by adopting the Neyman-Pearson lemma. Furthermore, in [4], the authors proposed an algorithm that samples outliers in a latent vector space to artificially generate anomalous machine sounds, in order to increase the difference of hypothesis test results between normal and anomalous sounds. Suefusa found that reconstruction errors are considerably decreased by interpolating only the center frame, rather than reconstructing the all input frames [5]. By minimizing the interpolation error, which is the difference between the output of an interpolation DNN(IDNN) and the input center frame, the performance of their ASD system was remarkably improved in the case of non-stationary machine sounds.

Task2 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 challenge [2] featured unsupervised detection of anomalous sounds from various machine conditions. The task provided a freely accessible machine sound database[6]-[7], which consists of both normal and anomalous operating sounds of six types of toy and real machines. Since only normal machine sounds are included in the training set, unsupervised learning-based approach was required for the task.

In this paper, we propose a novel ASD system using machine-adaptive focusing techniques. The principal contributions of this work are threefold. First, we used spectrograms as features instead of log-mel energies for more effective distinction between normal and anomalous machine sounds. Second, focused back-propagation and machine-adaptive focused hypothesis tests were applied in the training and evaluation phases, respectively. Only selective frames were used when calculating reconstruction errors to improve system performance in the unsupervised learning scenario. Third, we validated the suggested spectrogram features and focusing techniques on the baseline AE and IDNN models, and achieved remarkable improvement of the area under the receiver operating characteristic (ROC) curve (AUC) scores with the provided dataset.

2. REVIEW OF THE BASELINE SYSTEM

The AE-based baseline system [2] of DCASE Challenge task2 consists of 9 fully-connected layers where batch normalization and rectified linear unit (ReLU) activation functions are applied for all layers except for the output layer. Each machine sound sequence is converted to the time-frequency domain using short-time Fourier transform (STFT) with frame size of 1024 and half overlap. After that, every 5 consecutive STFT coefficients are fed into a mel-filter bank to obtain 128 dimension log-mel spectra feature vector. The input feature vector $\mathbf{x}$ can be reconstructed as follows:

$$\hat{\mathbf{x}} = D\{E\{\mathbf{x}|\theta_E\}|\theta_D\} \quad (1)$$

This work was supported by the Korea Government ICT R&D program of MSIT/IITP[2020-0-00857, Development of Cloud Robot Intelligence Augmentation, Sharing and Framework Technology to Integrate and Enhance the Intelligence of Multiple Robots]

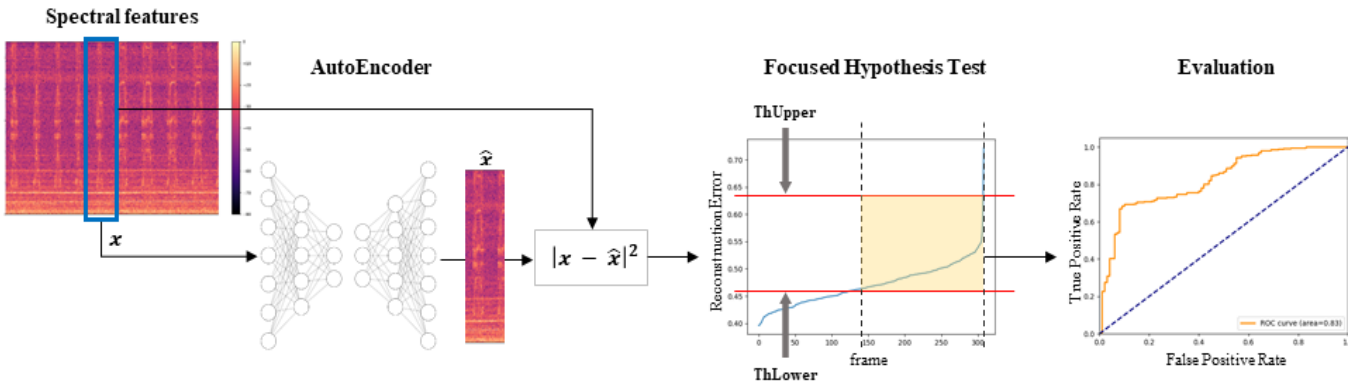

Figure 1: The block diagram of the proposed ASD system

Here, E and D denote the encoder and decoder parts of AE respectively, and θ_E and θ_D correspond to model parameters. Anomaly score $\mathcal{A}$ is defined as the MSE between input $\mathbf{x}$ and reconstructed vector $\hat{\mathbf{x}}$ as given by

$$\mathcal{A}(\mathbf{x}, \hat{\mathbf{x}}) = E\{||\mathbf{x} - \hat{\mathbf{x}}||_2^2\} \quad (2)$$

where $E\{\cdot\}$ and $||\cdot||_2$ denote mathematical expectation and L_2 norm, respectively. Finally, the machine sound is classified as abnormal when the anomaly score $\mathcal{A}$ exceeds threshold value.

3. THE PROPOSED ASD SYSTEM

In this section, we describe the proposed ASD system using several techniques including spectral feature extraction, focused back-propagation, and focused hypothesis test in order to improve ASD performance in the unsupervised learning scenario. The simplified block diagram of our proposed ASD system is depicted in Fig. 1. Each method is described in the following subsections.

3.1. Spectral feature extraction

In the provided baseline system, the AE model minimizes reconstruction error between input and reconstructed log-mel spectrogram feature vector during the training process. However, in [2], some types of machine sounds have innate characteristics that limit further improvement of ASD performance. To overcome such limitations, in [8], several time-frequency-domain spectral features [9]-[10] were extracted from the training set and their effects on the performance of the baseline ASD system were verified. In our system, we used the spectrogram feature as it showed the most significant performance improvement among various features. In Fig. 2, comparison of feature vectors extracted from a normal valve sound is shown. As shown (a) and (b) in Fig. 2, there are significant pattern losses when using the log-mel spectrogram feature vector. Since such dimension reduction may lead to performance degradation in ASD, submitted systems were trained on linear-scale-based spectral feature vectors to overcome loss of recognizable patterns.

3.2. Focused back-propagation

To improve the ASD performance, it is ideal to maximize the RE difference between normal and anomalous sounds. In the supervised learning scenario, different strategies can be considered for

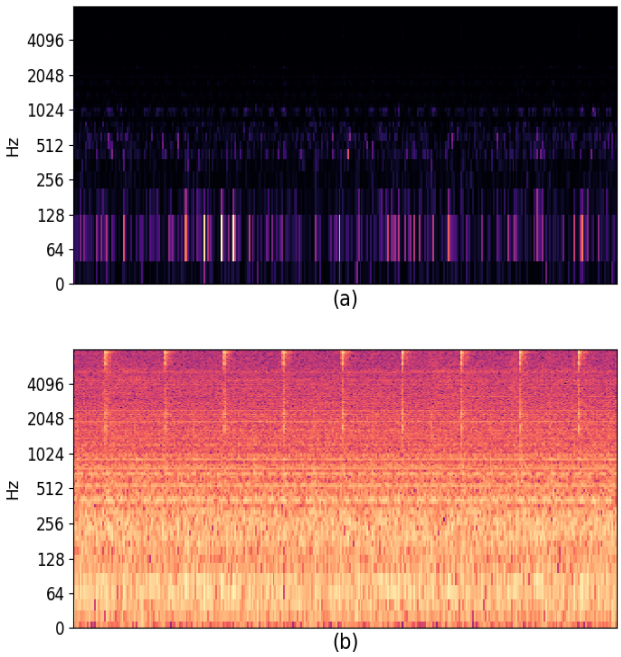

Figure 2: Comparison of spectral feature vectors extracted from normal value sound (a) log-mel spectrogram (b) spectrogram

the machine sound and the noise periods. However, the provided training set does not provide any label that can distinguish between machine and noise periods. In this work, focused back-propagation (FB) is proposed to further reduce RE in the normal machine sound periods in unsupervised learning scenarios. It is assumed that the energy of the input feature is greater in the machine sound periods than in the noise periods, and this assumption is identically applied to the reconstruction error. We select only the top p reconstruction error values in each batch so that only focused errors are used for back-propagation. In an unsupervised learning scenario, the RE of model with FB and baseline model are compared between normal and abnormal machine sound to see if using FB actually can enlarge the RE of the actual machine sound period. The comparison results are represented in Fig. 3. As shown in Fig. 3 (c) and (f), the RE of

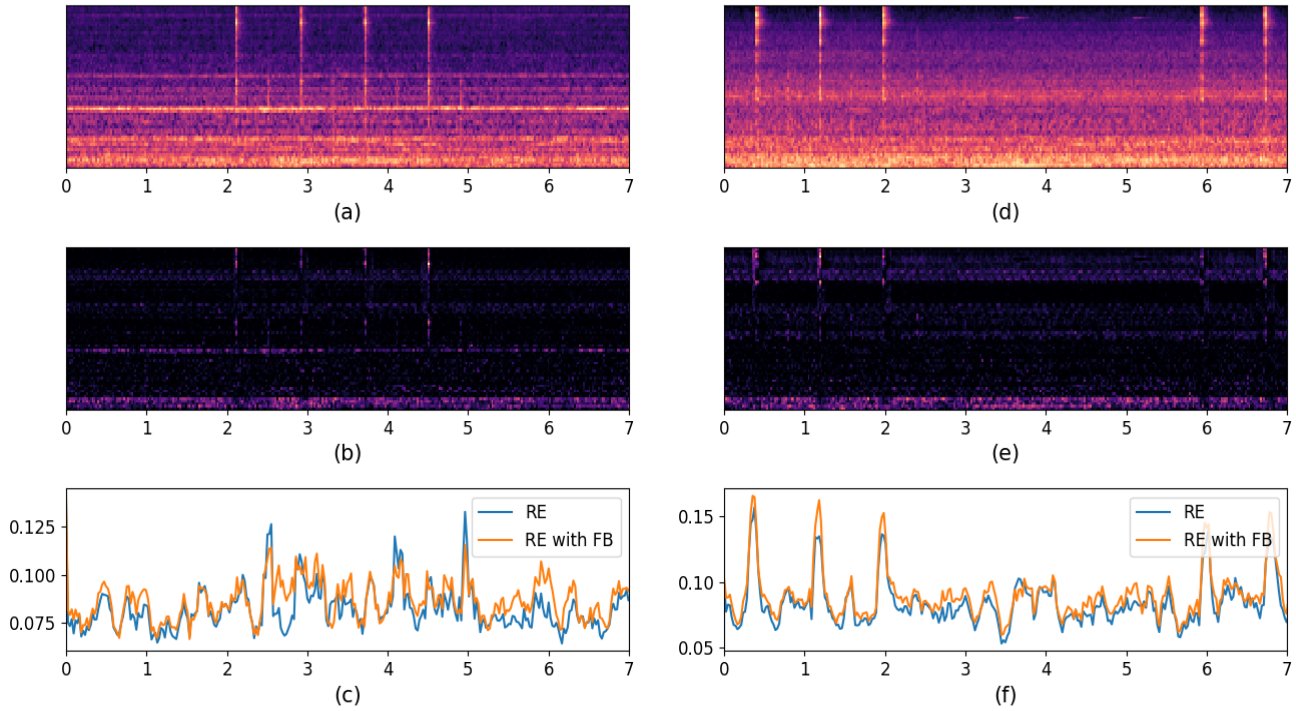

Figure 3: Comparison of RE results for normal and anomalous valve machine sound input signal (a) spectrogram of normal sound (b) reconstruction error of normal sound (c) comparison of RE results of normal sound (d) spectrogram of anomalous sound (e) reconstruction error of anomalous sound (f) comparison of RE results of anomalous sound

the FB-based AE makes the RE deviation larger in actual machine sound periods.

3.3. Focused hypothesis test

After the cost function of AE model converges, baseline ASD system makes decision according to hypothesis test results $\mathcal{H}$ as follows:

$$\mathcal{H}(\mathcal{A}(\mathbf{x}, \hat{\mathbf{x}})) = \begin{cases} 1(\text{Anomalous}), & \mathcal{A}(\mathbf{x}, \hat{\mathbf{x}}) > \phi \\ 0(\text{normal}), & \text{otherwise} \end{cases} \quad (3)$$

where ϕ is a pre-defined threshold value. Since averaged RE in frames are considered when deciding the status of machine sound in (3), focusing on the RE of only the machine sound periods affects the entire ASD performance. However, in the unsupervised learning scenario, it is difficult to precisely select the actual machine sound period. As an indirect alternative to overcome this problem, machine-adaptive focused hypothesis test was conducted in the evaluation phase of our submitted system. We found that RE of machine sound periods is lower than noise period for machine types valve and slider in case of which the frame energies of normal machine sounds are significantly larger than noise sounds. On the other hand, other types of machine sound show an opposite pattern. From these observations, we rectify the RE of frames by sorting them in ascending order first, then rejecting outliers as given by

$$\tilde{\mathcal{A}}(\mathbf{x}, \hat{\mathbf{x}}, \phi_l, \phi_u) = \begin{cases} \text{pass}, & \phi_l < \mathcal{A}_{sort}(\mathbf{x}, \hat{\mathbf{x}}) < \phi_u \\ \text{reject}, & \text{otherwise} \end{cases} \quad (4)$$

where $\mathcal{A}_{sort}(\mathbf{x}, \hat{\mathbf{x}})$ is a sorted version of $\mathcal{A}(\mathbf{x}, \hat{\mathbf{x}})$ in ascending order. Additionally, ϕ_l and ϕ_u are the threshold values for focusing RE, which are chosen empirically and differently for each machine. The concept of the focused hypothesis test is depicted in Fig. 1.

4. EXPERIMENTS AND SUBMISSIONS

We evaluated our system performances using the officially provided training set [6]-[7]. The training set consists of only normal machine sounds. Each machine sound was recorded with a single microphone and was sampled at 16 kHz. The recorded machine sound contains the normal machine sound as well as the factory noise signal, and label of the machine sound periods was not provided. To train our ASD systems, The spectrogram feature was extracted with a frame size of 1024 and half overlap by using numerical python library, librosa [11]. For optimal learning, ADAM optimizer with learning rate of 0.001 was set to training models of our ASD system. AUC and pAUC scores of baseline AE [2] and IDNN [5] models were compared to determine the effect of the proposed FB and FHT on performance. The encoder network comprises FC(D_{in}, 64, ReLU and BN), FC(64, 64, ReLU and BN), FC(64, 64, ReLU and BN), and FC(64, 8, ReLU and BN); the decoder network comprises FC(8, 64, ReLU and BN), FC(64, 64, ReLU and BN), FC(64, 64,

Table 1: Comparison of performances of FB and FHT applied to each model

	AUC (%)						pAUC (%)					
Methods	1	2	3	4	5	6	1	2	3	4	5	6
Baseline [2]	77.16	69.08	65.60	71.61	85.33	65.00	66.87	57.86	52.79	60.95	68.03	50.25
+ FB	82.11	71.96	66.69	71.04	86.36	69.31	72.90	60.95	52.99	62.18	68.46	51.04
+ FHT	77.57	73.13	68.13	**72.98**	90.21	80.43	67.49	60.50	54.30	60.76	73.90	54.00
+ FB and FHT	**82.56**	**74.27**	**68.81**	71.94	**90.29**	**85.29**	**73.25**	**74.27**	**68.81**	**71.94**	**90.29**	**85.29**
Baseline (spectrogram)	76.01	70.77	63.49	74.58	92.60	82.88	54.77	59.95	51.57	62.93	77.22	55.67
+ FB	80.53	69.06	62.40	74.32	92.08	76.82	66.39	58.84	50.99	61.03	75.99	53.44
+ FHT	79.30	**72.11**	**64.65**	**75.69**	**94.35**	**90.83**	63.16	**61.73**	**52.36**	**63.65**	**84.87**	**74.09**
+ FB and FHT	**81.84**	70.78	63.40	75.13	94.13	88.18	**68.28**	60.51	51.39	62.54	83.92	68.35
IDNN [5]	78.98	74.43	70.13	73.14	86.99	87.72	**72.03**	60.69	53.58	61.87	67.91	66.55
+ FB	80.04	72.96	68.85	72.35	85.89	87.60	71.58	60.81	53.68	62.23	68.43	67.81
+ FHT	79.41	**76.24**	**72.09**	**75.82**	**92.30**	**99.12**	71.57	**63.66**	54.57	**61.93**	76.23	96.27
+ FB and FHT	**80.43**	74.58	70.04	72.92	92.19	99.06	70.99	63.32	**55.21**	60.72	**77.30**	**96.34**
IDNN (spectrogram)	77.05	71.75	63.73	70.25	93.09	94.73	66.08	59.60	51.56	60.39	78.22	82.84
+ FB	79.93	70.17	62.01	70.20	92.86	92.95	69.37	59.28	51.25	60.09	77.37	78.32
+ FHT	77.89	**72.00**	**64.51**	**74.29**	**96.11**	**99.41**	66.95	**60.92**	**51.92**	**64.37**	**88.10**	**97.71**
+ FB and FHT	**80.36**	70.83	62.43	73.66	96.05	99.29	**69.50**	60.69	51.48	62.57	87.14	97.25

ReLU and BN), and FC(64, D_{out}, none), where FC(a, b, f) represents a fully-connected layer with input dimension a, an output demension b, and activation function f, respectively. For logmel and spectrogram features, (64×5, 64×5) and (513×5, 513×5) were used as the input and output dimensions (D_{in}, D_{out}) of the AE. In case of the IDNN model, (64×4, 64×1) and (513×4, 513×1) were accordingly used as input and output dimensions. Comparison of performances of FB and FHT applied to each model are summarized in Table 1 where machine classes are replaced as numbers as follows: ToyCar(1), ToyConveyor(2), fan(3), pump(4), slider(5), valve(6). Back-propagation was performed with only the top 10% errors in FB. Performance for each machine was repeatedly measured up to 10 times. In the baseline model with logmel feature, both FB and FHT improved, and the use of both FB and FHT showed the best performance. In particular, the performance improvement was largest for the valve machine sound. For the IDNN and baseline models with spectrogram, there were performance improvements in only the pAUC score for FB. In contrast, performance improvement was observed for all models when FHT was applied. As for the ToyCar machine sound, AUC and pAUC scores of all models significantly increased when using both FB and FHT. Compared to other machines, ToyCar is difficult to identify the machine sound and noise periods using the feature, but the performance was greatly improved by FB and FHT methods.

5. SUBMITTED ASD SYSTEMS

We applied FB and FHT to the AE model using 128-dimensional log-mel and 513-dimensional spectrogram, and reported it to task2 challenge [8]. Table 2 shows the performance of the submitted system for development set and evaluation set, respectively. As shown in the table, the proposed FB and FHT methods influenced the performance improvement.

Table 2: Performance results of reported ASD system in DCASE 2020 Task2

Machine	Development set		Evaluation set	
	AUC (%)	pAUC (%)	AUC (%)	pAUC (%)
ToyCar	82.73	70.35	81.40	66.37
ToyConveyor	76.61	64.01	86.41	71.92
Fan	76.21	62.06	84.38	64.23
Pump	70.77	54.50	82.30	59.97
Slider	94.16	83.97	96.39	83.58
Valve	89.67	72.85	83.86	61.99
Average	**81.69**	**67.96**	**85.79**	**68.01**

6. CONCLUSIONS

In this paper, we proposed focusing techniques including focused back-propagation and focused hypothesis test. In addition, we compared the performances of the baseline AE and IDNN models using spectral features, and two focusing methods of FB and FHT. Through experiments, we verified that these methods are effective in improving the detection performance of anomalous machine sounds in an unsupervised learning scenario.

7. REFERENCES

[1] T. Tagawa, Y. Tadokoro, and T. Yairi, "Structured denoising autoencoder for fault detection and analysis," in Proc. 6th Asian Conference on Machine Learning, 2015, pp. 96–111.

[2] Y. Koizumi, Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and Discussion on DCASE2020 Challenge Task2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring," 2020. [Online]. Available: arXiv:2006.05822.

[3] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in Proc. 25th European Signal Processing Conference (EUSIPCO), 2017, pp. 698-702.

[4] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212-224, Jan. 2019.

[5] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp.271-275

[6] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection," 2019. [Online]. Available: arXiv:1909.09347.

[7] Y. Koizumi, A. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 313-317.

[8] J. Park and S. Yoo, "Unsupervised detection of anomalous machine sound using various spectral features and focused hypothesis test in the reverberant and noisy environment," DCASE2020 Challenge, 2020.

[9] J. Driedger, M. Müller, and S. Disch, "Extending Harmonic-Percussive Separation of Audio Signals," in Proc. the international symposium on music information retrieval (ISMIR), 2014, pp. 611-616.

[10] A. Buades, B. Coll, and J. M. Morel, "A non-local algorithm for image denoising," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 60-65.

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in Proc. the 14th python in science conference, 2015, pp. 18-25.