APPLICATION OF SPECTRO-TEMPORAL RECEPTIVE FIELD ON SOFT LABELED SOUND EVENT DETECTION

Technical Report

Deokki Min, Hyeonuk Nam, Yong-Hwa Park

Korea Advanced Institute of Science and Technology
Department of Mechanical Engineering, 291 Daehak-ro
Yuseong-gu, Daejeon 34141, South Korea
{minducky, frednam, yhpark}@kaist.ac.kr

ABSTRACT

Spectro-Temporal Receptive Field (STRF) is a linear function which describe the relationship between sound stimulus and primary auditory cortex (A1) neural response within human auditory system. By means of convolution with sound spectrogram and STRF, we could simulate the A1 cell response which reacts to spectral and temporal modulation information of sound. By applying STRF, we expect SED model to detect sound events as human auditory system does. In this work, we used STRF as a kernel in convolutional layer and construct the two-branch deep learning model. One branch using STRF extracts the neuroscience-inspired spectro-temporal modulation information. The other is vanilla CNN branch which extracts complementary time-frequency information of input spectrogram those are missed by the STRF branch. TB-STRFNet, the proposed two-branch model, outperforms the baseline by 6.8% in terms of F1 score with optimum threshold per class.

Index Terms— Sound event detection, auditory system, STRF, spectral and temporal modulation

1. INTRODUCTION

Sound event detection (SED) is a task for recognition of occurring polyphonic sound events and their respective timing [1-4]. As human percepts multiple sound events spontaneously, possibility for development of sound event detection may lie on the characteristic of human auditory system. However, the characteristic of auditory system is yet to be totally comprehensible. As a role of each component inside human auditory system is entangled with other components, detailed mechanism behind the whole system are unclear. To understand the opaque characteristic of auditory system, many studies are still ongoing [5-7].

One of the efforts to understand auditory system is spectro-temporal receptive field (STRF) [8]. Neurons in auditory system represent various stimulus dependent properties [9]. STRF of a neuron reflects those spectral and temporal properties of sound stimulus that influence the firing probabilities of specific neuron [10]. Eggermont *et al.* [11] has revealed that neuron response can be predicted by convolution of STRF and spectrogram which represents both spectral and temporal properties of sound simultaneously. Primary auditory cortex (A1) is first relay station for auditory information and makes sense of information which is processed at multiple preceding stages in auditory system [6]. Higher hierarchy information such as sound event is also processed in A1, and it helps human to percept multiple sound events spontaneously. A1 neuron response can also be predicted by convolution of A1 STRF and sound stimulus.

STRF, which is used to predict A1 response, can be either estimated or constructed. Estimation methods include reverse correlation [11], boosting [12] and machine learning method such as support vector machine (SVM) [13]. On the other hand, STRF can be constructed by observing physiological data. Chi *et al.* [14] proposed STRF construction method by considering that A1 neurons are very sensitive to dynamic modulation of sound which is very important for speech intelligibility. Constructed STRF by Chi *et al.* [14] is used to various tasks requiring reproduction of human auditory system [15-17]. Especially, Vuong *et al.* [15] used STRF as kernel of deep learning model's convolutional layer. Recently, various deep learning models and methods are proposed for better performance of SED, and they aim to resemble working principle of human auditory perception behaviors [4, 18-20]. However, method using STRF as a kernel of convolutional layer has not been yet utilized in SED.

In this work, we adopted STRF construction method by Chi *et al.* [14] and STRF kernel method by Vuong *et al.* [15] to apply STRF on SED. We proposed two-branch model, TB-STRFNet, in which one branch extracts spectro-temporal modulation of sound input using STRF and the other branch extracts complementary time-frequency information with vanilla convolution module. TB-STRFNet outperformed baseline by 6.8% in terms of F1 score with optimum threshold per class [21].

2. METHODS

2.1. STRFConv

A convolutional layer whose kernel is constructed STRF is named as STRFConv in [15]. STRF filter is constructed with two parameters which are scale (Ω) and rate (ω), respectively. STRF examples which change along with scale and rate are shown in Figure 1.

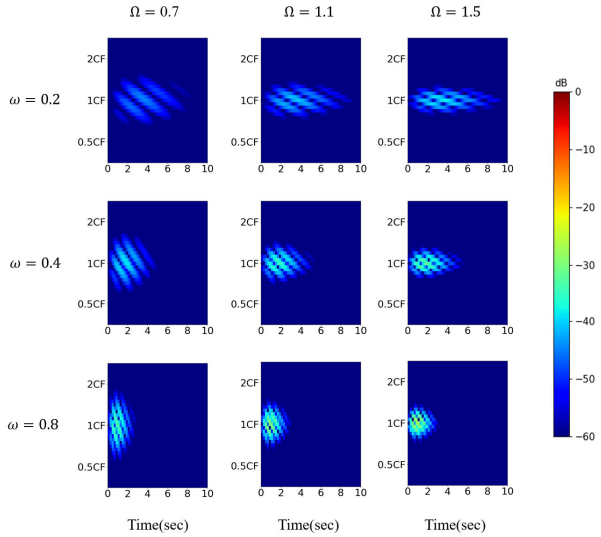

Figure 1: Examples of STRF kernels in STRFConv.

Axes of STRF represents time, and frequency range for x-axis, and y-axis, respectively. Given that STRF is centered at its center frequency (CF) [14], STRF is centered at its 1CF. Additionally, as observed in physiological data [22, 23], STRF frequency range is more than about 2 octaves which is from 0.5CF to 2CF. From first column to third column, scale gradually increases while rate is fixed. As scale increases, spectral spacing of ripple gets narrower while temporal spacing is constant. While high scale STRF represents spectrally narrow-tuned neuron characteristic, low scale STRF represents spectrally broad-tuned neuron characteristic. In contrast, from first row to third row, rate gradually increases while scale is fixed. As rate increases, temporal spacing of ripple gets narrower while spectral spacing is constant. While high rate STRF represents impulse-reactive neuron characteristic, low rate STRF represents prolonged-duration reactive property of neuron.

STRF is constructed with scale and rate as parameters to mimic spectro-temporal variously-tuned neuron characteristic. STRFConv utilizes various STRF as a kernel of convolutional layer to reflect those variously-tuned neuron characteristic inside auditory system, and it use scale and rate as learnable parameters to adjust STRF for given task.

2.2. TB-STRFNet

Proposed TB-STRFNet is an SED model which consists of two branches. The model structure of TB-STRFNet is depicted in Figure 2. Input for each branch is the same mel-spectrogram. Both branches are composed of a convolutional layer followed by six convolution blocks. One branch adopts its first layer by STRFConv while the other adopts vanilla convolution. 64 STRF kernels are used in STRFConv. whereas half of kernels are down direction STRF which capture decreasing spectral modulation as time passes and the other half are upward direction STRF which capture increasing spectral modulation as time passes. Downward and upward direction STRF is shown in Figure 3(a) and 3(b), respectively. The following convolution blocks consist of 2d convolutional layer, batch normalization, ReLU activation and 2d maxpool layer.

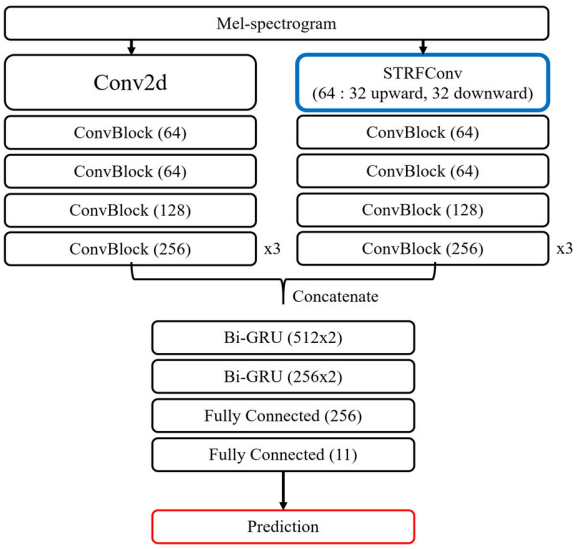

Figure 2: A structure of TB-STRFNet.

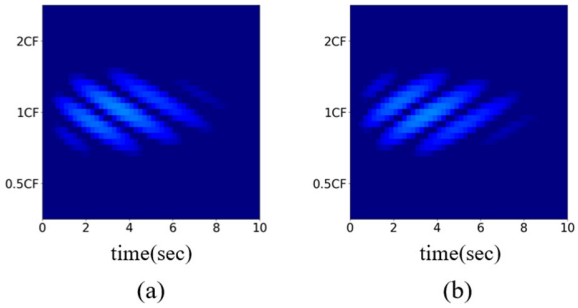

Figure 3: (a) Downward direction STRF (b) Upward direction STRF.

The STRF branch, which includes STFRConv, extracts neuroscience-inspired modulation information by its STRF kernel so that SED model could extract sound event information in a way similar to human auditory system. On the other hand, the Vanilla branch composed of only vanilla convolutional layers extracts complementary sound event information that might be missed by STRF branch. Extracted feature map from two branches are concatenated and go through remaining layers which are composed of two Bi-GRU layers and two fully connected layers.

2.3. Implementation Details

Mel-spectrogram is used as input feature, by 44.1kHz sampling rate, 8,820 hop size, 17,640 window length and 64 mel-bin. For training, epoch number is 150, batch size is 32, mean-square error

Table 1: A test result for submission systems.

Model	Params	ER(%)	$F1_m$(%)	$F1_M$(%)	$F1_{MO}$(%) (Main)
Baseline	0.38M	48.18	70.79	35.79	42.91
System 1	3.56M	**44.50**	**72.78**	**36.12**	**45.81**
System 2	3.56M	44.70	72.53	35.2	45.37
System 3	3.85M	46.10	70.20	27.82	45.41
System 4	3.85M	45.30	71.00	29.83	44.27

for loss function and Adam optimizer are used. 5 cross-fold validation setup is used for stable overall evaluation. Soft labeled data are used for training [24], whereas hard labeled data are used for test. All evaluation metrics are segment-based calculated by 1-sec time resolution. Micro-average error rate (ER), micro-average F1 score ($F1_m$), macro-average F1 score ($F1_M$) and macro-average F1 score with optimum threshold per class ($F1_{MO}$) are used for evaluation metrics [21].

3. RESULTS

Total four systems are submitted for evaluation DCASE Task 4 subtask B. A test result for each submission is represented in Table 1. 5 cross-fold validation is used that 5 best models are created for each session. Test result is based on created 5 best models.

System 1 is single-best TB-STRFNet without ensemble. System 1 result is only based on one session in which 5 best models are created due to 5 cross-fold validation. System 1 performs the best performance with 6.8% increase compared to the baseline about $F1_{MO}$ which is the main metric of the task. System 2 is TB-STRFNet with ensemble which utilizes six sessions that is total thirty models. System 2 used average decision-making method for ensemble. System 3 is TB-STRFNet with global embeddings of AST, which utilizes external dataset. It is only based on one session that no ensemble is used. The size of global embeddings of AST is 256. Embeddings are concatenated with extracted feature from basic branch and STRF branch in TB-STRFNet. System 4 is TB-STRFNet with global embeddings of AST and ensemble. Averaging is also used for decision-making method while ensemble of system 4.

4. REFERENCES

[1] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound Event Detection: A tutorial," *IEEE Signal Processing Magazine,* vol. 38, pp. 67-83, September 01, 2021

[2] T. Virtanen, M. D. Plumbley, and D. Ellis, "Computational analysis of sound scenes and events." Springer, 2018.

[3] J. Ebbers and R. Haeb-Umbach, "Pre-Training And Self-Training For Sound Event Detection In Domestic Environments," Paderborn University, Tech. Rep, 2022.

[4] H. Nam et al., "Heavily Augmented Sound Event Detection utilizing Weak Predictions." arXiv preprint arXiv:2107.03649.

[5] M. S. Malmierca, "Auditory system," in The rat nervous system: Elsevier, 2015, pp. 865-946.

[6] J. Schnupp, I. Nelken, and A. King, "Auditory neuroscience: Making sense of sound." MIT press, 2011.

[7] K. Jasmin, C. F. Lima, and S. K. Scott, "Understanding rostral–caudal auditory cortex contributions to auditory perception," Nature Reviews Neuroscience, vol. 20, no. 7, pp. 425-434, 2019.

[8] A. Aertsen and P. I. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grassfrog: I. Characterization of tonal and natural stimuli," *Biological Cybernetics,* vol. 38, no. 4, pp. 223-234, 1980.

[9] J. J. Eggermont, "Context dependence of spectro-temporal receptive fields with implications for neural coding," Hearing research, vol. 271, no. 1-2, pp. 123-132, 2011.

[10] A. M. H. J. Aertsen and P. I. M. Johannesma, "The Spectro-Temporal Receptive Field," Biological Cybernetics, vol. 42, no. 2, pp. 133-143, 1981/11/01 1981.

[11] J. J. Eggermont, A. M. H. J. Aertsen, and P. I. M. Johannesma, "Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field," *Hearing Research,* vol. 10, no. 2, pp. 167-190, 1983/05/01/ 1983.

[12] S. V. David, N. Mesgarani, and S. A. Shamma, "Estimating sparse spectro-temporal receptive fields with natural stimuli," Network: Computation in neural systems, vol. 18, no. 3, pp. 191-212, 2007.

[13] A. F. Meyer, M. F. Happel, F. W. Ohl, and J. Anemüller, "Estimation of spectro-temporal receptive fields based on linear support vector machine classification," BMC Neuroscience, vol. 10, pp. 1-2, 2009.

[14] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," The Journal of the Acoustical Society of America, vol. 106, no. 5, pp. 2719-2732, 1999/11/01 1999, doi: 10.1121/1.428100.

[15] T. Vuong, Y. Xia, and R. Stern, "Learnable Spectro-temporal Receptive Fields for Robust Voice Type Discrimination," p. arXiv:2010.09151.

[16] R. Sharma, T. Vuong, M. Lindsey, H. Dhamyal, R. Singh, and B. Raj, "Self-supervision and Learnable STRFs for Age, Emotion, and Country Prediction," p. arXiv:2206.12568.

[17] C.-Y. Wang et al., "Spectral–temporal receptive field-based descriptors and hierarchical cascade deep belief network for guitar playing technique classification," IEEE Transactions on Cybernetics, vol. 52, no. 5, pp. 3684-3695, 2020.

[18] H. Nam, S.-H. Kim, B.-Y. Ko, and Y.-H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," p. arXiv:2203.15296.

[19] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: IEEE, pp. 4308-4312.

[20] G.-T. Lee, H. Nam, S.-H. Kim, S.-M. Choi, Y. Kim, and Y.-H. Park, "Deep learning based cough detection camera using enhanced features," Expert Systems with Applications, vol. 206, p. 117811, 2022.

[21] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold Independent Evaluation of Sound Event Detection Scores," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: IEEE, pp. 1021-1025.

[22] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra," Journal of neurophysiology, vol. 76, no. 5, pp. 3503-3523, 1996.
[23] R. C. DeCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," science, vol. 280, no. 5368, pp. 1439-1444, 1998.
[24] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.