

Data Augmentation and Squeeze-and-Excitation Network on Multiple Dimension for Sound Event Localization and Detection in Real Scenes

Technical Report

Byeong-Yun Ko, Hyeonuk Nam, Seong-Hu Kim, Deokki Min, Seung-Deok Choi, Yong-Hwa Park

Korea Advanced Institute of Science and Technology
Department of Mechanical Engineering, 291 Daehak-ro,
Yuseong-gu, Daejeon 34141, South Korea

{b.y.ko, frednam, seonghu.kim, minducky, haroldchoi6, yhpark}@kaist.ac.kr

ABSTRACT

Performance of sound event localization and detection (SELD) in real scenes is limited by small size of SELD dataset, due to difficulty in obtaining sufficient amount of realistic multi-channel audio data recordings with accurate label. We used two main strategies to solve problems arising from the small real SELD dataset. First, we applied various data augmentation methods on all data dimensions: channel, frequency and time. We also propose original data augmentation method named *Moderate Mixup* in order to simulate situations where noise floor or interfering events exist. Second, we applied Squeeze-and-Excitation block on channel and frequency dimensions to efficiently extract feature characteristics. Result of our trained models on the STARSS22 test dataset achieved the best ER, F1, LE, and LR of 0.53, 49.8%, 16.0°, and 56.2% respectively.

Index Terms— Sound Event Localization and Detection, Data augmentation, Squeeze-and-Excitation network

1. INTRODUCTION

Sound event localization and detection (SELD) is a task detecting desired sound events with their class, timestamps (onset and offset of the sound event), and direction-of-arrival (DoA) [1, 2, 3]. Its application includes surveillance, robotics, monitoring system, etc. Before deep learning based SELD methods, spatial information of sound source was usually predicted by rule-based methods using multi-channel audio signals obtained from microphone array formats such as conventional microphone beamformer, differential and super-directive array, circular/spherical microphone array, binaural microphone array, etc. [4, 5]. However, sound signals obtained in real scenes are distorted by bias error by array platform, reverberation, and background/interference noise. In addition, moving sources degrades the performance of SELD system due to difficulty in recognizing DoA from constantly changing room impulse response (RIR) of sound source. These problems have been limiting performance of rule-based DoA estimation methods [6]. To overcome such problems arising from unpredictable acoustic characteristics in real scenes, data driven methods have been suggested. Especially, deep learning-based methods achieved excellent performance robust to noise and moving source by learning complex pattern of sound source from SELD dataset synthesized

by RIRs and target sound event samples [1, 2, 3]. However, to implement SELD on real environments robustly, it should be trained and tested on real SELD dataset, which is recorded on real spatial sound scenes. But obtaining sufficient amount of real SELD data is time consuming and requires extensive works to consider various real environments and precisely label the timestamps and DoA of sound events. Although synthesized dataset has complemented scarce real dataset, dataset characteristics such as sound event occurrences, sound event densities, and amount of polyphony differ from real dataset. In this context, previous deep learning-based methods developed with synthesized dataset tend to be overfitted to synthesized dataset.

We used two main strategies to overcome these problems: 1) applying data augmentation techniques in a way it addresses all dimensions of multi-channel audio data: channel, frequency and time and 2) applying Squeeze-and-Excitation (SE) block on channel and frequency dimension. Data augmentations were applied to Sony-Tau Realistic Spatial Soundscapes 2022 (STARSS22) spatial recording [7] and DCASE2022 synthetic SELD mixtures [8] datasets to prevent overfitting and reduce difference between real and synthesized dataset by transforming acoustic characteristics of sound signal. We used following data augmentation methods: channel swap for channel dimension, pitch shift for frequency dimension, frame shift and time masking for time dimension. In addition, we applied an original data augmentation method proposed in this work, *Moderate Mixup* which further enhance SELD task performance. In addition, we applied SE block on channel and frequency dimensions to apply optimal weighting of channels and frequency components on extracted CNN outputs.

2. PROPOSED METHODS

2.1. Implementation Details

For input feature, we used spatial cue-augmented log spectrogram (SALSA) consisting of 4 channel log-linear spectrograms and 3 channel eigenvector-based intensity vector [9] on FOA microphone array because it gives additional directional information by approximation of steering vector with eigen principal component analysis. The log-linear spectrogram with 200 frequency bins was extracted using 512 window length and 300 hop length from an

Table 1: A comparison of SELD performance on different output formats. Best score for each metric in each output format is in bold, where the best among all format is in red bold

Output format	SED threshold	ER	F1(%)	LE(°)	LR(%)
Decoupled output [9]	0.3	0.75	22.3	38.8	40.9
	0.5	0.70	21.8	40.4	35.1
	0.7	0.72	21.5	42.3	32.0
ACCDOA [10]	0.3	0.93	25.2	37.2	56.1
	0.5	0.64	30.4	31.4	41.8
	0.7	0.63	26.2	54.8	28.7
Multi-ACCDOA [11]	0.3	0.91	22.5	41.9	58.0
	0.5	0.68	27.8	37.5	44.4
	0.7	0.65	27.0	48.0	31.1

audio sample with 24 kHz sampling rate. The sound events and source localizations were recognized using convolutional recurrent neural network (CRNN) architecture with ResNet22 and 2 BiGRU layers [9]. Models presented in this work is trained using both STARSS22 [7] and DCASE2022 synthetic SELD mixtures [8]. For evaluation, we used DCASE2022 Challenge Task3 evaluation metrics [1] consisting of four metrics: location-dependent error rate (ER) and F1 score, class-dependent localization error (LE), and localization recall (LR) in the following analysis. For location-dependent ER and F1 score, a sound event is considered correctly detected only if the class prediction is correct and localization error is less than 20°.

2.2. SELD Output Formats

We compared SELD performance on STARSS22 test dataset depending on output formats: decoupled SELD output (performing event detection and localization separately) [1, 9], ACCDOA [10], and multi-ACCDOA [11] in Table 1. For each format, we applied sound event detection (SED) threshold value of 0.3, 0.5 and 0.7. SED threshold is a criterion that determines if an event is active or not, by applying the threshold value on SED prediction for decoupled SELD output and length of Cartesian DoA vector for ACCDOA and multi-ACCDOA. We observed that optimal SELD performance is achieved on SED threshold of 0.3 for decoupled SELD output and 0.5 for ACCDOA and multi-ACCDOA respectively, considering the number of best performance metrics on each output format. While multi-ACCDOA is supposed to perform better than ACCDOA by increased prediction capability on detection of overlapping events from the same class [11], it performed worse. This seems to be because of larger output dimension of multi-ACCDOA to consider overlapping sound events from same class, which induced underfitting to the small real recording dataset. Based on these results, we applied ACCDOA on the output prediction format with SED threshold of 0.5.

2.3. Data Augmentation Methods on Multiple Dimensions

As mentioned in 2.1, we used external data (DCASE2022 synthetic SELD mixtures [8]), generated by convolving target sound events from FSD50K [12] with measured room impulse responses of Eigenmike in various acoustic scenes [13], to supplement sparse real recording data of STARSS22. Synthetic dataset which consists of 1200 sound samples, is much larger than STARSS22 with 67 sound samples. To prevent overfitting to much larger synthetic SELD dataset, various data augmentation methods were applied to reduce the difference between the synthetic sounds and

Table 2: A comparison of SELD task performance depending on various data augmentations and combinations. CS, PS, FS, TM, and MM refers to channel swap, pitch shift, frame shift, time masking, and moderate mixup respectively. Best score for each metric in each data augmentation method is in bold, where the best among all data augmentation method is in red bold

Methods	ER	F1(%)	LE(°)	LR(%)
Baseline (CS with probability of 0.5 + PS with frequency shift range of 10)	0.61	39.8	20.1	53.3
FS with probability of 0.3	0.62	36.1	19.8	52.1
FS with probability of 0.5	0.61	41.0	20.0	57.8
FS with probability of 0.7	0.63	38.8	23.2	58.2
FS with probability of 0.9	0.59	39.4	25.5	51.2
TM with mask ratio from 1/6 to 1/10	0.61	39.3	19.7	52.5
TM with mask ratio from 1/8 to 1/15	0.60	37.1	19.8	55.2
TM with mask ratio from 1/10 to 1/20	0.62	40.1	19.0	55.0
TM with mask ratio from 1/12 to 1/25	0.61	38.1	20.6	52.1
MM with probability of 0.3	0.60	38.2	22.0	55.8
MM with probability of 0.5	0.58	42.6	18.9	57.4
MM with probability of 0.7	0.61	43.2	22.8	52.5
MM with probability of 0.9	0.59	40.2	20.9	50.6
best FS + best TM	0.60	42.2	22.0	55.3
best FS + best MM	0.57	43.6	19.2	58.8
best TM + best MM	0.58	43.8	18.4	55.6
bset FS + best TM + best MM	0.61	42.8	20.9	54.1

the real recording dataset by transforming acoustic characteristics of sound signal. We applied data augmentation methods to SALSA on FOA microphone array verified in [9] which are channel swap (CS) [14] and pitch shift (PS, which is the same as frequency shift in [9]). These two data augmentation methods make transformations on channel and frequency dimension of input feature respectively. Dimension of microphone channel is important as it gives physical information regarding the direction of sound events [4]. Frequency dimension affects performance of SED when appropriately handled [15, 16, 17]. In addition, time dimension should also be addressed for SELD because temporal information is crucial in audio domain [18, 19, 20, 21] which handles time-series data. Thus, we applied following data augmentation methods which are shown to enhance SED performance: frame shift (FS) which shifts the input feature and label along time axis [15], and time masking (TM) which masks the random time region of input feature and label [22]. Time masking involves a parameter ‘mask ratio’ randomly drawn from mask ratio range (set as hyperparameter).

In addition, we applied originally proposed data augmentation method named moderate mixup. Moderate mixup is ACCDOA format-based data augmentation method designed to simulate environments where interferences and background noises contaminate target sound events. In ACCDOA format, an active sound event is expressed by a unit vector pointing towards the location of the sound source. Applying mixup [23] directly on ACCDOA unit vectors for two separate sound events with same class would result in label pointing toward direction between the locations of sound sources corresponding to the sound event, where no sound source actually would present. To avoid such distortion of DoA, moderate mixup mixes input features with a mixing ratio selected from beta distribution so that mixing ratio is close to 0 and 1 mostly and the direction information will be represented by the dominant sound source with mixing ratio closer to 1 by taking ACCDOA unit vector of the dominant sound source. As a result, by only mixing input feature and not mixing label. Mixed data simulates interferences and background noises.

Table 3: A comparison of SELD task performance depending on network models. SE, CS, PS, FS, TM, and MM stands for SE block, pitch shift, frame shift, time masking, and moderate mixup respectively.

Methods	ER	F1(%)	LE(°)	LR(%)
no SE with CS + PS	0.61	39.8	20.1	53.3
Original SE with CS + PS	0.58	40.5	19.6	52.1
Original SE with CS + PS + FS + MM	0.56	43.6	17.1	59.1
Original SE with CS + PS + TM + MM	0.58	40.6	17.5	51.7
Multi-dimensional SE with CS + PS	0.59	41.1	19.0	55.8
Multi-dimensional SE with CS + PS + FS + MM	0.59	42.3	18.5	58.3
Multi-dimensional SE with CS + PS + TM + MM	0.58	44.5	17.9	58.4

We experimented on the data augmentation methods at time dimension and moderate mixup to validate each method for SELD and figure out optimal combinations of data augmentation methods. The results are shown in Table 2. Baseline refers to CRNN base-line model with ACCDOA as illustrated in 2.1. The baseline includes channel swap (probability of 0.5) and pitch shift (frequency shift range of 10) as default, as they are already proven in [9]. It is shown that frame shift and time masking, data augmentation methods on time dimension, enhance F1, LE, and LR. In addition, moderate mixup largely enhances F1, and LE from baseline. For the experiment on combinations of data augmentation methods, we observed that applying one of time dimension data augmentation methods with moderate mixup results in best performances, while applying both results in performance degradation from excessive data transformation.

2.4. Squeeze-and-Excitation block

There are 13 sound event classes to detect on DCASE 2022 Task3, including various sound events from harmonic sounds (e.g. human speech and music) to impulse-like sounds (e.g. clapping and door open/close). In addition, the effect of moving source changes acoustic characteristics of sound source constantly. Such harsh conditions make it hard to conduct event detection and predict localization information. Therefore, SELD system should accurately extract features of each sound events from varying complex feature maps. Therefore, we applied Squeeze-and-Excitation (SE) block [24] on ResNet22 structure to adaptively adjust channel weights depending on convolution input with minimal increase of model size. In addition, we applied SE block on frequency dimension as well on each time frame of input to adaptively adjust frequency weights as well.

We compared performances of trained models, and the results are described in Table 3. From 2nd to 4th rows, original SE block which applies on channel dimension is applied. From 5th to 7th rows, multi-dimensional SE block which applied SE on frequency dimension and then on channel dimension is applied. It is shown that applying original SE block on network enhances ER, F1, and LE. Moreover, when utilizing additional data augmentation methods, frame shift and moderate mixup, to the model, LR increases to 59.1. We also observed that applying time masking and moderate mixup with multi-dimension SE block, F1 score largely increases compared to applying frame shift and moderate mixup on the model.

Table 4: Final results of the models submitted.

Models	ER	F1(%)	LE(°)	LR(%)
DCASE2022 Task3 baseline [25]	0.71	21%	29.3	46
(no SE with CS + PS + FS + MM)×2				
(Original SE with CS + PS + FS + MM)×3	0.53	49.8	16.0	55.9
(Multi-dimensional SE with CS + PS + FS + MM)×3				
(Multi-dimensional SE with CS + PS + FS + MM)×3	0.55	46.2	16.4	54.6
(Multi-dimensional SE with CS + PS + FS + MM)×3	0.57	46.4	17.2	54.4
(Multi-dimensional SE with CS + PS + TM + MM)×3	0.55	46.4	17.0	56.2

3. RESULTS

The results of submitted models for DCASE 2022 Task3 on the STARS22 test dataset are shown in Table 4. Baseline in Table 4 refers to DCASE 2022 Challenge Task3 baseline, different from baseline in Table 2. Model 1 is ensemble average of the 11 models trained with all technics proposed in this paper. Model 2 is an ensemble average on the 3 models trained with original SE block and all data augmentation except time masking. Model 3 is ensemble average on the 3 models trained with multi-dimensional SE block and all data augmentation except time masking. Model 4 is an ensemble average on 3 models trained with multi-dimensional SE block and all data augmentation except frame shift.

4. ACKNOWLEDGMENT

This work was supported by ‘‘Human Resources Program in Energy Technology’’ of the Korea Institute of Energy Technology Evaluation and Planning (KETEP), granted financial resource from the Ministry of Trade, Industry & Energy, Republic of Korea. (No. 20204030200050)

5. REFERENCES

- [1] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, ‘‘Overview and evaluation of sound event localization and detection in DCASE 2019,’’ *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 684-698, 2020.
- [2] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, ‘‘Joint measurement of localization and detection of sound events,’’ in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 333-337.
- [3] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumley, ‘‘Polyphonic sound event detection and localization using a two-stage strategy,’’ in *Proc. Detection Classification Scenes Events Workshop*, 2019, pp.30-34.
- [4] Y. H. Kim and J. W. Choi, *Sound Visualization and Manipulation*, Singapore: John Wiley & Sons Singapore Pte. Ltd., 2013.
- [5] G. T. Lee, S. M. Choi, B. Y. Ko, and Y. H. Park, ‘‘HRTF measurement for accurate identification of binaural sound localization cues,’’ arXiv preprint arXiv:2203.03166, 2022.

- [6] I. J. Jung and J. G. Ih, "Compensation of inherent bias errors in using the three-dimensional acoustic intensimetry for sound source localization," *J. Sound Vib.*, vol. 461, Nov. 2019.
- [7] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," arXiv preprint arXiv:2206.01948, 2022.
- [8] DCASE2022 Task 3 Synthetic SELD mixtures for baseline training. [Online]. Available: <https://doi.org/10.5281/zenodo.6406873>.
- [9] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "SALSA: Spatial Cue-Augmented Log-Spectrogram Features for Polyphonic Sound Event Localization and Detection," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1749-1762, 2022.
- [10] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *Proc. of IEEE ICASSP*, 2021, pp. 915-919.
- [11] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training," in *Proc. of IEEE ICASSP*, 2021, pp. 316-320.
- [12] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 829-852, 2021.
- [13] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A Dataset of Dynamic Reverberant Sound Scenes with Directional Interferers for Sound Event Localization and Detection," arXiv preprint arXiv:2106.06999, 2021.
- [14] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First Order Ambisonics Domain Spatial Augmentation for DNN-based Direction of Arrival Estimation," arXiv preprint arXiv:1910.04388, 2019.
- [15] H. Nam, B. Y. Ko, G. T. Lee, S. H. Kim, W. H. Jung, S. M. Choi, and Y. H. Park, "Heavily augmented sound event detection utilizing weak predictions," arXiv preprint arXiv:2107.03649, 2021.
- [16] H. Nam, S.-H. Kim, and Y.-H. Park, "FilterAugment: An Acoustic Environmental Data Augmentation Method," in *Proc. of IEEE ICASSP*, 2022, pp. 4308-4312.
- [17] H. Nam, S. H. Kim, B. Y. Ko, and Y. H. Park, "Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection," arXiv preprint arXiv:2203.15296, 2022.
- [18] 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 Syst. Appl.*, vol. 206, Nov. 2022.
- [19] S. H. Kim and Y. H. Park, "Adaptive Convolutional Neural Network for Text-Independent Speaker Recognition," in *Proc. Interspeech 2021*, pp. 641-645.
- [20] S. H. Kim, H. Nam, and Y. H. Park, "Temporal Dynamic Convolutional Neural Network for Text-Independent Speaker Verification and Phonemic Analysis," in *Proc. of IEEE ICASSP*, 2022, pp. 6742-6746.
- [21] S. H. Kim, H. Nam, and Y. H. Park, "Decomposed Temporal Dynamic CNN: Efficient Time-Adaptive Network for Text-Independent Speaker Verification Explained with Speaker Activation Map," arXiv preprint arXiv:2203.15277, 2022.
- [22] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.0877, 2019.
- [23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," arXiv preprint arXiv:1710.09412, 2017.
- [24] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," in *Proc. of IEEE CVPR*, 2018, pp. 7132-7141.
- [25] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 34-48, Mar. 2019.