SOUND EVENT LOCALIZATION AND DETECTION USING SQUEEZE-EXCITATION RESIDUAL CNNS

Technical Report

Javier Naranjo-Alcazar[1,2]*, Sergi Perez-Castanos*[2]*, Maximo Cobos*[2]*, Francesc J. Ferri*[2]*, Pedro Zuccarello*[1]

[1] Instituto Tecnológico de Informática, València, Spain {jnarnajo, pzuccarello}@iti.es
[2] Universitat de València, Burjassot, Spain, {pecaser@alumni.uv.es, {maximo.cobos, francesc.ferri}@uv.es}

ABSTRACT

Sound event localisation and detection (SELD) is a problem in the field of automatic listening that aims at the temporal detection and localisation (direction of arrival estimation) of sound events within an audio clip, usually of long duration. Due to the amount of data present in the datasets related to this problem, solutions based on deep learning have positioned themselves at the top of the state of the art. Most solutions are based on 2D representations of the audio (different spectrograms) that are processed by a convolutional-recurrent network. The motivation of this submission is to study the squeeze-excitation technique in the convolutional part of the network and how it improves the performance of the system. This study is based on the one carried out by the same team last year. This year, it has been decided to study how this technique improves each of the datasets (last year only the MIC dataset was studied). This modification shows an improvement in the performance of the system compared to the baseline using MIC dataset.

***Index Terms*—** SELD, Deep Learning, Convolutional Recurrent Neural Network, Squeeze-Excitation, Residual learning, DCASE2021

1. INTRODUCTION

Sound event localisation and detection (SELD) corresponds to the machine listening problem that aims to detect and localise a sound event in an audio clip of typically long duration [1, 2, 3]. The detection consists in correctly classifying the sound event into one of the predefined classes while setting the time at which the event starts and ends. For several editions of DCASE, this task only attempted to solve the SED problem. On the other hand, localisation consists in estimating the direction of arrival (DOA) of the source producing the event in terms of azimuth and elevation angles. SELD proposes a joint problem involving these two areas, for which a single system capable of performing both detection and localisation must be proposed. For an intelligent system to be able to estimate directional information, audio signals must have been recorded by a set of microphones (multi-channel audio input).

The task of Sound Event Location and Detection (SELD) has been constantly in evolution in the scope of the DCASE Challenge until reaching the problem presented in this edition. In 2013 [4], the problem to be solved was known as Sound Event Detection (SED). In the 2016 [5] and 2017 [6] editions, this problem was again proposed as a task. In this case, the objective was to create a system capable of detecting the onset and offset of sound events while correctly classifying to which class these events belong. The first time that event localisation was raised in addition to the SED problem was in the 2019 edition [7, 8]. Last year, 2020 [9], the dataset was modified in addition to the metrics (setting a threshold of 20° in the detection metrics). This edition incorporates a new consideration which is the existence of directional interferences, meaning sound events out of the target classes that are also point-like in nature. This is a much more accurate recreation of a real environment [10].

The work done in this edition can be seen as a extension of the work done in the last edition [11]. Last year, an analysis was made of how squeeze-excitation and residual techniques [12, 13] (applied in the convolutional part of the system) can lead to a more robust system without any extra modification of the baseline. For this, the *Conv-StandardPOST* block was implemented and it was analysed how the different ratios (ρ) influenced the system. Furthermore, it was compared with a residual block that did not implement any squeeze-excitation technique, please see Fig. 3. According to the results obtained in the Challenge, the block with $\rho = 1$ obtained the best position. So, since this study was performed only using the MIC dataset, this year we will analyse how the *Conv-StandardPOST* block with a fixed ratio behaves with each of the datasets.

This technical report is organized as follows: Section 2 introduces the network presented as the baseline and the modification done in this work to achieve the improvement. Section 3 explains the dataset used and the training procedure. Section 4 shows the results obtained by the framework implemented and Section 5 concludes our work.

2. METHOD

2.1. Baseline System

The baseline network is known as SELDnet [7]. The main modification this year has been the elimination of the SED classification branch, adopting a joint training (ACCDOA) that unites SED loss and DOA in a homogenous regression vector loss [14].

The first module of the system is the feature extractor. It obtains multi-channel audio representations. With the MIC dataset, 10 channels (GCC) are obtained and with the first-order ambisonics (FOA), 7 channels (Intensity vector). These representations were introduced in [2].

2.2. Squeeze-Excitation Residual blocks and modifications to the baseline network

This submission is understood as a extension of the work done in [13, 11] where different different squeeze-excitation modules where studied [15] plus the contribution of two novel blocks using

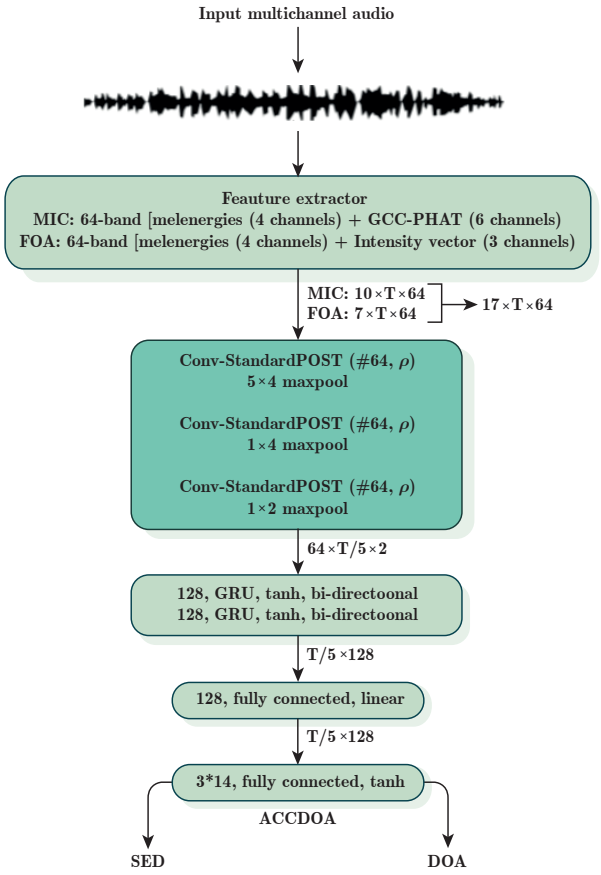

Figure 1: SELD framework proposed in this work. The most highlighted block corresponds to the change made in this task. The lighter blocks have the same configuration as in the baseline. ρ indicates the ratio parameter. In this work, $\rho = 1$.

the *Concurrent Spatial and Channel Squeeze and Channel Excitation* (scSE) configuration presented in [12]. The implementation of the scSE block is detailed in Figure 2. Following the conclusions of [13] and [11], in the present work, the convolutional layers of SELDnet are replaced by the *Conv-StandardPOST* blocks with $\rho = 1$. The number of filters remains the same (64 filters). The framework proposed in this work is shown in Figure 1.

The code used for this experimentation can be found in the following link[1].

3. EXPERIMENTAL DETAILS

3.1. Dataset

The dataset used in this edition is the one defined as TAU-NIGENS Spatial Sound Events 2021. The major change this year is the addition of sound events that do not belong to the target classes. For a more detailed description of the dataset, visit the following link[2] and the paper [9].

Concerning the usage of the samples (see Table 1), in the development phase, three folders are used for training, one for validation and one for testing. In this stage, the ground truth of all the samples is available. However, in the evaluation stage, 4 folders are used for training, 1 for validation and 2 for testing. In this case, the ground truth of the test samples is not available (Challenge results).

Stage	training	validation	test
Development	3-6	2	1
Evaluation	2-6	1	7-8

Table 1: Distribution of the folders in the two stages. Each folder contains 100 samples.

3.2. Training procedure

The training process is the same as that proposed in the baseline [7, 9]. However, it has been decided to implement a learning rate decay system. Thus, if the performance of the system is not improved within 15 epochs, the learning rate decreases by a factor of 0.5. The training is terminated if there is no improvement in 30 epochs.

[1] https://github.com/Machine-Listeners-Valencia/seld-dcase2021

[2] http://dcase.community/challenge2021/task-sound-event-localization-and-detection

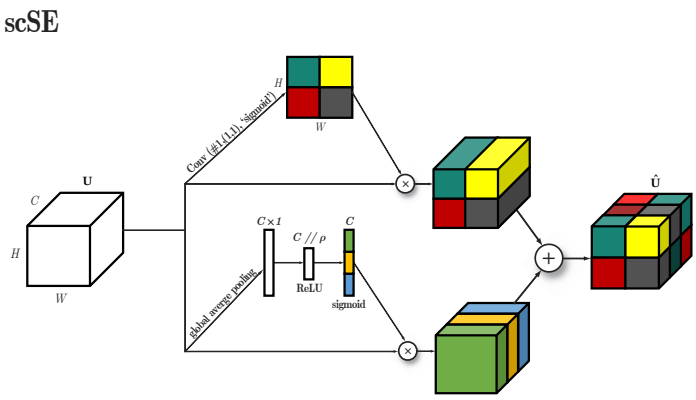

Figure 2: scSE composed by Spatial Squeeze-Excitation (sSE) module (top branch) and channel Squeeze-Excitation (cSE) module (lower branch) [13, 12]

4. RESULTS

In order to study the squeeze-excitation residual blocks in both datasets, it was decided to carry out the experimentation with $\rho = 1$ as it showed the best performance in last year submission [11].

4.1. Metrics

The metrics used in this task are known as location-dependent. The detection of an event will be considered correct if the angle prediction is below a threshold set at 20°.

In this task, there are 2 metrics per output. Two metrics to measure the robustness of the detection are the error rate (ER_{20^o}) and the F-score (F_{20^o}) with a threshold of 20º. On the other hand, the localisation accuracy is measured with the localisation error (LE_{CD}) and the localisation recall (LR_{CD}). For more insight about metrics, please see [10].

4.2. Development stage

The results obtained in both datasets independently and combined (concatenating the representations obtained with each one) are shown below (see Table 2). If the FOA dataset is considered, it can be seen that the metric that is improved is LR_{CD}. However, LE_{CD} decreases by 4º. On the other hand, the MIC dataset is greatly improved, the system shows a better performance in all metrics. The most improved metric is F_{20^o}, improving by 5.2 percentage points. Finally, if both representations are concatenated, no considerable improvement can be seen with respect to the FOA dataset.

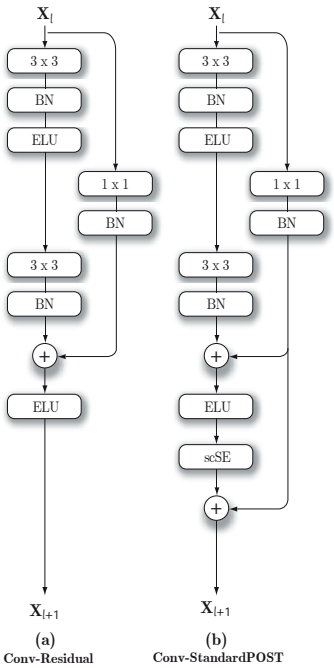

Figure 3: Residual blocks analyzed in this paper. BN stands for Batch Normalization and scSE for squeeze-excitation module. Convolutional layers are indicated with the kernel size.

Framework	*Dataset*	ER_{20^o}	F_{20^o}	LE_{CD}	LR_{CD}
Baseline	FOA	0.69	33.9%	24.1º	43.9%
Proposed	FOA	0.71	31.9%	27.6º	**46.6%**
Baseline	MIC	0.74	24.7%	30.9º	38.2%
Proposed	MIC	**0.72**	**30.2%**	**29.4º**	**42.5%**
Proposed	FOA-MIC	0.71	31.3%	27.9º	46.7%

Table 2: Accuracy (%) results obtained compare with the proposed baseline

5. CONCLUSION

The motivation for this work is the study of squeeze-excitation techniques for the improvement of SED/DOA systems. For this purpose, it has been decided to modify only the convolutional part of the system and to follow the conclusions obtained in last year's edition. Despite the interferences present in this edition, it can be observed that, using the MIC dataset, all the metrics are improved. However, we cannot observe the same behaviour using the FOA dataset. This suggests further study if this dataset is to be used.

6. ACKNOWLEDGEMENTS

The participation of Dr. Cobos and Dr. Ferri is supported by ERDF and the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-097045-B-C21, as well as grants AICO/2020/154 and AEST/2020/012 from Generalitat Valenciana.

7. REFERENCES

[1] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," DCASE2019 Challenge, Tech. Rep., June 2019.

[2] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," DCASE2019 Challenge, Tech. Rep., June 2019.

[3] W. Xue, T. Ying, Z. Chao, and D. Guohong, "Multi-beam and multi-task learning for joint sound event detection and localization," DCASE2019 Challenge, Tech. Rep., June 2019.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct 2015.

[5] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb 2018.

[6] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, in press.

[7] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[8] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and uetection," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019. [Online]. Available: http://dcase.community/documents/challenge2019/technical_reports/DCASE2019_Adavanne.pdf

[9] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[10] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://https://arxiv.org/abs/2009.02792

[11] J. Naranjo-Alcazar, S. Perez-Castanos, J. Ferrandis, P. Zuccarello, and M. Cobos, "Task 3 dcase 2020: Sound event localization and detection using residual squeeze-excitation cnns," DCASE2020 Challenge, Tech. Rep., July 2020.

[12] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

[13] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112 287–112 296, 2020.

[14] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, June 2021.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00745