TASK 3 DCASE 2020: SOUND EVENT LOCALIZATION AND DETECTION USING RESIDUAL SQUEEZE-EXCITATION CNNS

Technical Report

Javier Naranjo-Alcazar[1,2], *Sergi Perez-Castanos*[1], *Jose Ferrandis*[1], *Pedro Zuccarello*[1], *Maximo Cobos*[2]

[1] Visualfy, Benisano, Spain, {javier.naranjo, sergi.perez, jose.ferrandis, pedro.zuccarello}@visualfy.com,
[2] Universitat de València, Burjassot, Spain, {janal2}@alumni.uv.es, {maximo.cobos}@uv.es

ABSTRACT

Sound Event Localization and Detection (SELD) is a problem related to the field of machine listening whose objective is to recognize individual sound events, detect their temporal activity, and estimate their spatial location. Thanks to the emergence of more hard-labeled audio datasets, Deep Learning techniques have become state-of-the-art solutions. The most common ones are those that implement a convolutional recurrent network (CRNN) having previously transformed the audio signal into multichannel 2D representation. In the context of this problem, the input to the network, usually, has many more channels than in other problems related to machine listening. This is because the audio is recorded by an array of microphones.Some frequency representation is obtained for each of them together with some additional representations, such as the generalized cross-correlation (GCC), whose objective is the assessment of the relationship between channels. This work aims to improve the accuracy results of the baseline CRNN by adding residual squeeze-excitation (SE) blocks in the convolutional part of the CRNN. The followed procedure involves a grid search of the parameter *ratio* of the residual SE block, whereas the hyperparameters of the network remain the same as in the baseline. Experiments show that by simply introducing the residual SE blocks, the results obtained in the development phase clearly exceed the baseline.

Index Terms— SELD, Deep Learning, Convolutional Recurrent Neural Network, Squeeze-Excitation, Residual learning, DCASE2020

1. INTRODUCTION

Sound Event Localization and Detection (SELD) tries to solve both problems, related to machine listening, of tracking the activation of different classes (detection) and the spatial localization of sound events at the same time [1, 2, 3, 4]. For an intelligent system to be able to calculate such outputs, the audio must have been recorded by an array of microphones (multichannel audio input).

SELD first appeared in DCASE 2019 edition as an evolution of the Sound Event Detection (SED) problem. SED was presented in the first edition of the DCASE in 2013 [5] and was presented again as a task in the 2016 [6] and 2017 [7] editions. The objective of this task is the individual detection of particular events that occur in a scene. The nature of this problem is directly confronted with the polyphonic nature of audio [8, 9], i.e. the overlapping of several events in the same time period. SELD task DCASE2020 edition can be seen as a modification from 2019 DCASE challenge. Modifications done in this edition have been the presented dataset, that has been increased, and the detection metrics that are computed with a 20^o threshold from the reference for true positives.

Regarding the dataset called TAU-NIGENS Spatial Sound Events 2020 [10], it should be observed that each scene has been recorded in two different formats: using an array of 4 microphones (MIC) and with first-order Ambisonics (FOA). In both recording formats (MIC or FOA), each sound event in the scene is associated with a direction-of-arrival (DoA) to the recording point, and temporal onset and offset times. The number of classes to be detected are 14. Some of these classes are: piano, male speech, female speech, barking dong, among others. As it can be noticed, sounds belonging to these classes are easily found in domestic environments. This encourages the proposal of solutions that could improve real-world applications such as home assistants [11].

For this submission, MIC recording format has been used. In the MIC setup, the microphones have been placed on an spherical acoustically-hard baffle, and their positions described in spherical coordinates, ϕ, θ and r are as follows:

- M1: (45^o, 35^o, 4.2cm)
- M2: (-45^o, -35^o, 4.2cm)
- M3: (135^o, -35^o, 4.2cm)
- M4: (-135^o, 35^o, 4.2cm)

Some of the modifications of the dataset presented in this edition with respect to the previous one are the following:

- (2x) Large lecture halls with inclined floor. Ventilation noise.
- (2x) Modern classrooms with multiple seating tables and carpet flooring. Ventilation noise.
- (2x) Meeting rooms with hard floor and partially glass walls. Ventilation noise.
- (2x) Old-style large classrooms with hard floor and rows of desks. Ventilation noise.
- Large open space in underground bomb shelter, with plastic floor and rock walls. Ventilation noise.
- Large open gym space. People using weights and gym equipment.

The dataset is divided into several folders under development. 4 folders (3-6) are used for training, folder 2 for validation and 1 for testing.

Regarding the difference in the metrics used in this edition, it is intended to have a more representative calculation of the problem by doing a joint evaluation of location and detection [12]. A prediction will be considered correct if both are of the same class and

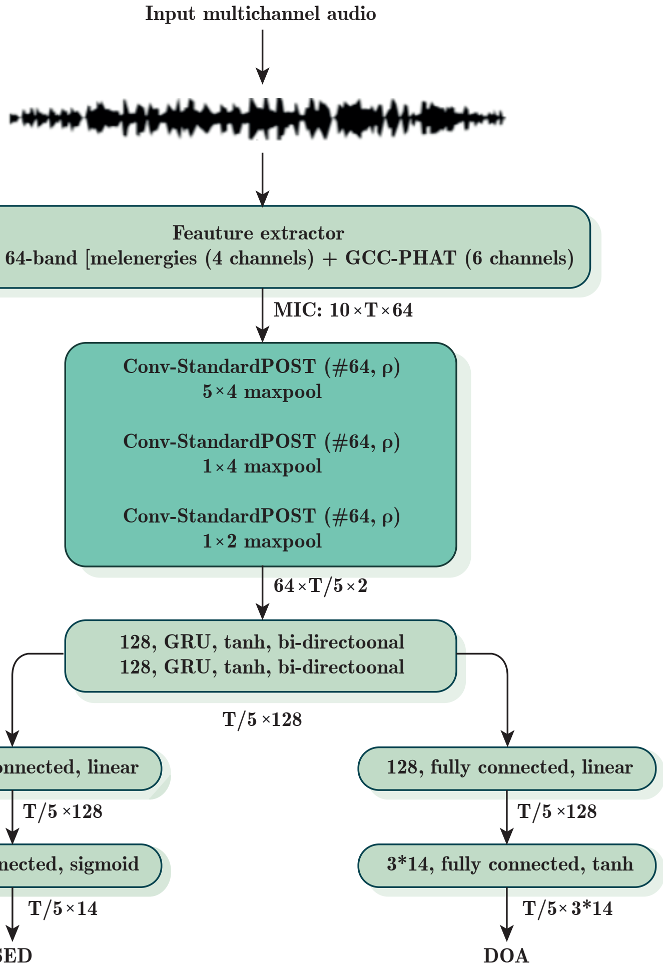

Input multichannel audio
Feauture extractor
MIC: 64-band [melenergies (4 channels) + GCC-PHAT (6 channels)
MIC: 10×T×64
Conv-StandardPOST (#64, ρ)
5×4 maxpool
Conv-StandardPOST (#64, ρ)
1×4 maxpool
Conv-StandardPOST (#64, ρ)
1×2 maxpool
64×T/5×2
128, GRU, tanh, bi-directoonal
128, GRU, tanh, bi-directoonal
T/5 ×128
128, fully connected, linear
T/5 ×128
14, fully connected, sigmoid
T/5×14
SED
128, fully connected, linear
T/5 ×128
3*14, fully connected, tanh
T/5×3*14
DOA

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.

the distance between them is below 20°. The detection metrics are now location-dependent. Regarding detection, the metrics use the error rate (ER_{20^o}) and the F-Score (F_{20^o}). On the other hand, in part of localization, the metrics used are the localization error (LE_{CD}), expressing average angular distance between predictions and references of the same class and the localization recall metric (LR_{CD}), expressing the true positive rate of how many of these location predictions were detected in a class, of the total occurrences of the class.

This paper aims to study the improvements that squeeze-excitation techniques can bring to the SED/DOA task. To this purpose, the convolutional part of the CRNN proposed as a baseline is modified. The convolutional layers are replaced by residual squeeze-excitation blocks. As it is the first time that these blocks are introduced in this problem, a grid-search of the hyperparameter ratio is performed. The results show that only by making this modification the results of the baseline are considerably exceeded.

This paper is organized as follows: Section 2 introduces the network presented as the baseline and the modification done in this submission to it. Section 3 shows the results obtained by the framework implemented and Section 4 concludes our work.

2. METHOD

2.1. Baseline System

The baseline network is known as SELDnet [13]. This network is a CRNN that uses detections (SED) to estimate the DOA of each of the classes. The SED is displayed as a multi-label classification and the DOA as a multi-output regression.

The network input is a representation of 10 channels (MIC format) of dimension $T \times F$, where T corresponds to the number of temporary bins and F to the number of frequency bins. In this case, F is set to 64 and T corresponds to 300 temporal bins. 4 of the channels correspond to the log-Mel Spectrograms of each signal recorded by each microphone of the array and the remaining 6 correspond to the calculation of the generalized cross-correlation (GCC) [2]. The implementation of the baseline network can be found in this link[1].

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

Most Machine Listening frameworks rely on the ability of the CNN to extract meaningful features. Either in a VGG-style [14] or Residual [15] networks are very similar between different submissions or proposed solutions. Therefore, the improvement of the systems often falls on other aspects such as data augmentation techniques (pitch shifting [16], speed perturbation [3] or mixup [17] among others) or the ensemble of many independent models [3, 18, 19, 20].

In [21] an analysis of different Residual Squeeze-Excitation blocks proposed in [22] plus the contribution of two novel blocks using the *Concurrent Spatial and Channel Squeeze and Channel Excitation* configuration presented in [23] is carried out in Acoustic Scene Classification task. The analysis is run without any data augmentation technique during training. In [21], a novel configuration labelled as *Conv-StandardPOST* showed the best results in treated ASC problem. Therefore, following the conclusions of [21], in the present work the convolutional layers of SELDnet are replaced by the *Conv-StandardPOST* blocks. The number of filters remain the same, 64. The framework proposed in this work is shown in Figure 1.

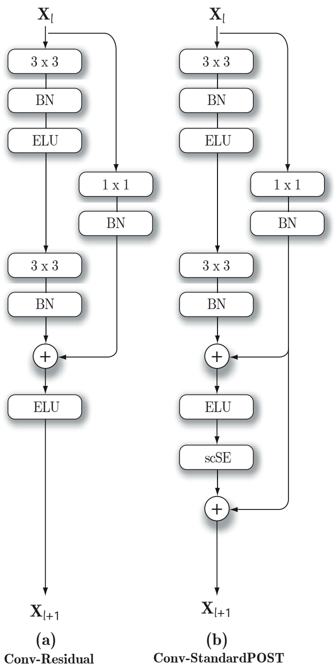
In order to widen the study of the contribution of the squeeze-excitation technique, the network was also trained with another residual configuration. The block labelled as *Conv-Residual* in [21] was also used in the present work. It is a residual block with certain particularities, but with no squeeze-excitation techniques. Both, *Conv-StandardPOST* and *Conv-Residual* configurations, can be seen in Figure 2. For further insight about this choice, see [21]. The code for this submission can be found in the following link[2]

2.3. Experimental details

The training process is the same as that proposed in the baseline. No hyperparameter, such as learning rate, the decay weight, number of epochs, etc., was modified; in this way, the variations in the results can only be attributed to the proposed modifications explained in section 2.2.

3. RESULTS

In order to study the squeeze-excitation residual blocks contribution, it was decided to carry out a grid search of different possible ratios. Keep in mind that the network is made up of 3 blocks of 64 filters. The ratio (ρ) is the same for all blocks as it can be seen in Figure 1. The results can be seen in Tables 1 and 2. It has also been decided to add a configuration that although it is residual does not incorporate the squeeze-excitation block in order to be able to independently analyze the contribution of residual learning and this same learning plus squeeze-excitation. The chosen block has been the *Conv-Residual* presented in [21]. The results are presented using the following structure: the system named *baseline* is the one presented as starting point by the organization, *Conv-Residual* corresponds to the residual block shown in Figure 2(a). Experiments indicated by ρ correspond to *Conv-StandardPOST* implementation with that particular ratio, see Figure 2(b).



(a) Conv-Residual

(b) Conv-StandardPOST

Figure 2: Residual blocks analyzed in this paper. Layers are indicated as Batch Normalization (BN), squeeze-extication module (scSE) and convolutional layers are indicated with the kernel size.

framework	ER	F (%)	LE (°)	LR (%)
baseline	*0.56*	*59.2*	*22.6*	*66.8*
Conv-Residual	0.50	**65.2**	19.0	68.5
$\rho = 1$	0.51	63.7	20.5	**69.1**
$\rho = 2$	0.52	62.2	19.4	68.1
$\rho = 4$	**0.49**	65.1	20.2	68.1
$\rho = 8$	0.51	64.0	19.4	67.4
$\rho = 16$	0.52	63.0	**18.6**	68.0

Table 1: Development results using DCASE2019 metrics (`'dev'`).

framework	$ER_{20°}$	$F_{20°}$ (%)	LE_{CD} (°)	LR_{CD} (%)
baseline	*0.78*	*31.4*	*27.3*	*59.0*
Conv-Residual	**0.68**	**42.3**	**22.5**	**65.1**
$\rho = 1$	0.70	39.2	23.5	63.6
$\rho = 2$	0.69	40.4	23.2	62.1
$\rho = 4$	**0.68**	40.9	23.3	65.0
$\rho = 8$	0.69	40.8	23.5	63.8
$\rho = 16$	0.69	40.7	23.3	62.8

Table 2: Development results using DCASE2020 metrics (`'dev'`).

[1] https://github.com/sharathadavanne/seld-dcase2020

[2] https://github.com/Joferesp/DCASE2020-Task3

As can be seen in Table 1, all the configurations exceed the metrics presented as baseline. Residual learning allows obtaining more accurate systems by adding only a shortcut, in our case convolutional (see *Conv-Residual* results). In turn, the concerned squeeze-excitation improves the results of residual learning without this process (*Conv-Residual*) in all metrics except F_{20^o}. However, there is no one ratio that exceeds the others, depending on the metric, a different ratio shows better performance.

The improvement provided by squeeze-excitation operations can be seen more clearly using the DCASE2019 metrics. This year's restrictions do not allow the improvement to be so marked (see Table 2). Although implementations with *Conv-StandardPOST* improve the proposed baseline, the architecture with *Conv-Residual* obtains the most accurate metrics. To continue with the study, the same tables are presented but in evaluation step, where 5 folders are used for training (2-6), 1 for validation (1) and 2 for testing (7-8). In this case the metrics are shown on the validation folder since the ground-truth of the test folders is not available.

framework	ER	F (%)	LE (^{o})	LR (%)
Conv-Residual	0.49	65.6	18.0	69.0
$\rho = 1$	**0.47**	**68.0**	**17.6**	71.1
$\rho = 2$	0.48	65.7	18.3	**71.7**
$\rho = 4$	0.48	66.2	19.1	71.6
$\rho = 8$	0.48	66.7	18.5	70.1
$\rho = 16$	0.48	66.9	17.8	**71.7**

Table 3: Evaluation results using DCASE2019 metrics ('eval').

framework	ER_{20^o}	F_{20^o} (%)	LE_{CD} (^{o})	LR_{CD} (%)
Conv-Residual *	0.64	45.8	20.7	65.4
$\rho = 1$ *	**0.63**	**47.0**	21.3	**67.9**
$\rho = 2$	0.65	44.9	**21.0**	65.5
$\rho = 4$	0.66	43.5	22.1	66.1
$\rho = 8$ *	0.64	46.0	21.7	66.6
$\rho = 16$ *	**0.63**	46.4	21.1	66.8

Table 4: Evaluation results using DCASE2020 metrics ('eval'). Implementations marked with * are those that have been submitted for the challenge.

As it can be appreciated in Tables 3 and 4, in evaluation step, worse results are obtained with the *Conv-Residual* block than in the development step. In fact, it is not the block that shows the best performance. This can lead us to two conclusions: the first is that squeeze-excitation techniques do contribute to more accurate training in SED/DOA task. The second is that these techniques require more data to achieve relationships that can be better generalized in the test step. With this data partition it can be argued that the implementation of *Conv-StandardPOST* block with $\rho = 1$ shows the best trade-off between SED and DOA tasks. Below is a Table with the name of the submission related to the implemented block to make it more understandable when analyzing the final results of the challenge.

Block used	Submission name
Conv-Residual	Naranjo-Alcazar_Vfy_task3_1
$\rho = 1$	Naranjo-Alcazar_Vfy_task3_2
$\rho = 8$	Naranjo-Alcazar_Vfy_task3_3
$\rho = 16$	Naranjo-Alcazar_Vfy_task3_4

Table 5: Relationship between the name of the submission and the implementation explained in this paper.

4. CONCLUSION

The aim of this paper is to analyze the improvements that residual learning and squeeze excitation techniques can bring in the field of SED and DOA. To this end, it has been decided to make as few modifications as possible to the framework presented as a baseline. By modifying only the convolutional part of it and without any extra technique during the learning (data augmentation) or during the inference phase (ensemble of several models) results that exceed the baseline to a greater extent have been achieved. However, the results show that this field is subject to further improvements and new solutions to enhance squeeze-excitation techniques in the SED and DOA fields.

5. ACKNOWLEDGEMENTS

The participation of Javier Naranjo-Alcazar and Dr. Pedro Zuccarello in this work is partially supported by Torres Quevedo fellowships DIN2018-009982 and PTQ-17-09106 respectively from the Spanish Ministry of Science, Innovation and Universities.

6. 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] J. Zhang, W. Ding, and L. He, "Data augmentation and prior knowledge-based regularization for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., June 2019.

[5] 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.

[6] 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.

[7] 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.

[8] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[9] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[10] 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

[11] S. Sigtia, A. M. Stark, S. Krstulović, and M. D. Plumbley, "Automatic environmental sound recognition: Performance versus computational cost," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2096–2107, 2016.

[12] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct 2019, accepted.

[13] 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

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] K. Noh, C. Jeong-Hwan, J. Dongyeop, and C. Joon-Hyuk, "Three-stage approach for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., June 2019.

[17] W. J. Jee, R. Mars, P. Pratik, S. Nagisetty, and C. S. Lim, "Sound event localization and detection using convolutional recurrent neural network," DCASE2019 Challenge, Tech. Rep., June 2019.

[18] T. N. T. Nguyen, D. L. Jones, R. Ranjan, S. Jayabalan, and W. S. Gan, "Dcase 2019 task 3: A two-step system for sound event localization and detection," DCASE2019 Challenge, Tech. Rep., June 2019.

[19] L. Mazzon, M. Yasuda, Y. Koizumi, and N. Harada, "Sound event localization and detection using foa domain spatial augmentation," DCASE2019 Challenge, Tech. Rep., June 2019.

[20] S. Leung and Y. Ren, "Spectrum combination and convolutional recurrent neural networks for joint localization and detection of sound events," DCASE2019 Challenge, Tech. Rep., June 2019.

[21] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *arXiv preprint arXiv:2003.09284*, 2020.

[22] 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

[23] 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.