TASK 1 DCASE 2020: ASC WITH MISMATCH DEVICES AND REDUCED SIZE MODEL USING RESIDUAL SQUEEZE-EXCITATION CNNS

Technical Report

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

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

ABSTRACT

Acoustic Scene Classification (ASC) is a problem related to the field of machine listening whose objective is to classify/tag an audio clip in a predefined label describing a scene location such as park, airport among others. Due to the emergence of more extensive audio datasets, solutions based on Deep Learning techniques have become the state-of-the-art. The most common choice are those that implement a convolutional neural network (CNN) having previously transformed the audio signal into a 2D representation. This two-dimensional audio representation is currently a subject of research. In addition, there are solutions that propose several concatenated 2D representations, thus creating a representation with several input channels. This article proposes two novel stereo audio representations to maximize the accuracy of an ASC framework. These representations correspond to the 3-channel representations such as the left channel, the right channel and the difference between channels $(L - R)$ using the Gammatone filter bank and the harmonic, percussive and difference between channels sources using the Mel filter bank. Both representations are also concatenated creating a 6-channel with different audio filter banks. Furthermore, the proposed CNN is a residual network that employs squeeze-excitation techniques in its residual blocks in a novel way to force the network to extract meaningful features from the audio representation. The proposed network is used in both subtasks with different modifications to meet the requirements of each one. However, since stereo audio is not available in Subtask A, the representations are slightly modified in that task. This technical report first presents the overlaps of the two tasks and then makes the relevant changes to each task in one section per task. The baselines are surpassed in both tasks by approximately 10 percentage points.

Index Terms— Deep Learning, Convolutional Neural Network, Acoustic Scene Classification, Mel-Spectrogram, Gammatone, HPSS, DCASE2020

1. INTRODUCTION

The analysis of daily sounds can be an improvement in different areas that are very prone to automation such as autonomous driving or surveillance. Acoustic scene classification (ASC) is the field of machine listening that aims to define the location of the scene (park, airport, etc.) based only on audio data [1, 2, 3, 4]. In turn, machine listening can be defined as the field of artificial intelligence whose final objective is to obtain information (location, classification, detection, etc.) intelligently from audio data. The illustration of an ASC framework can be found in Figure 1

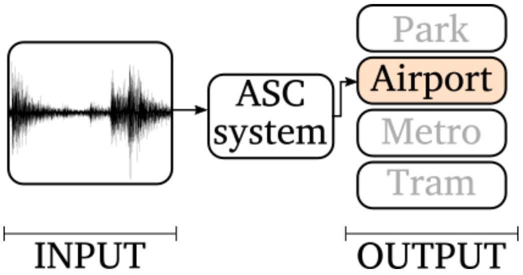


Figure 1: Acoustic Scene Classification framework. Given an audio input, the system must classify it into a given predefined class.

The objective of Task 1 of the DCASE 2020 Challenge is to encourage participants to propose different solutions to address the ASC problem in a public labeled audio dataset. In this edition, the ASC task is divided into two subtasks that introduce particular requirements to solve the ASC issue.

Subtask A proposes the problem of the ASC along with the concern that audio clips come from different audio sources. In fact, the audios have been recorded using 4 different recording devices. These devices are referred as the following:

A. Soundman OKM II Klassik / studio A3, electret binaural microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24 bit resolution

B. Samsung Galaxy S7

C. IPhone SE

D. GoPro Hero5 Session

In addition to these four devices, 11 simulated mobile devices are created using the audio collect by Device A. These artificial mobile devices are name S_i being i the number of the device $1 \leq i \leq 11$. Audios from S_i are calculated passing audio from device A through convolution with the selected S_i impulse response, then processed with a selected set of parameters for dynamic range compression (device specific). The full dataset are composed by 10 different scenes such as airport, tram, park, etc. and it has been recorded from 12 different cities like Barcelona, Helsinki or Lisbon among others. This dataset is named TAU Urban Acoustic Scenes 2020 Mobile [5]

For the development stage, a portion of this dataset is released. Audios that come from devices, A, B, C and $S_1 - S_6$ are released. Furthermore, only audios from 10 cities are released instead of the 12 available in the full dataset. The dasatet is splitted in a 70%/30%

training/validation configuration making a total of 13965 segments for training and 2970 segments for validation. The audios last 10 seconds and are provided in single channel 44.1kHz 24-bit format.

In evaluation stage, data from new devices will be released such as device D and $S_7 - S_{11}$ as well as from devices that have appeared in development step. The data in this set comes from the 12 possible cities. The final goal of this task is to create systems robust enough to different devices.

Subtask B proposes a slim version of an ASC problem where only three major classes must be targeted: indoor, outdoor and transportation. This goal must be accomplished with low complexity solutions in terms of model size.

The maximum allowed size is 500 KB for non-zero parameters. This corresponds to a model with 128000 parameters for *float32* precision. The layers whose objective is the extraction of features or said process carried out offline (for example the calculation of the log-Mel Spectrogram) does not count when adding parameters. However, other transfer learning techniques that use another previously trained network for feature extraction do mean an increase in the number of parameters because the parameters of the transfer learning network (VGGish [6], L3net [7], Soundnet [8], among others) must be added to the final classification network.

The dataset used in this subtask is named TAU Urban Acoustic Scenes 2020 3Class. As in Subtask A, only audios of 10 cities are available in development stage, the dataset having been recorded in 12 cities in total. In this case, only 3 major classes want to be classified: *indoor scenes, outdoor scenes* and *transportation related scenes*. Audios have been recorded with device A and are provided in 48kHz-24 bit format.

The objectives of this work are the following: firstly, the proposal of a new audio stereo representation using the Gammatone filter bank. Secondly, the proposal of a 6-channel representation whose channels are audio representations with different scales such as Mel and Gammatone. Finally, the analysis of the contributions of the residual and squeeze-excitation techniques in ASC with the restrictions or problems proposed in the subtasks.

The rest of this report is organized as follows: Section 2 explains the audio representations used in this submission plus the general configurations of our models. Section 3 details the specific modifications of the models for Subtask A. Section 4 explains the networks submitted in order to accomplish this subtask requirements. Both Section 3 and Section 4 show the obtained results in development stage comparing them with the proposed baseline. Finally, Section 5 concludes our work.

2. METHOD

2.1. Audio representaion

Following the idea of last year submission [9] a multi-channel 2D audio representation is used when possible. Harmonic (H) and percussive (P) [10, 11] log-Mel audio representation sources showed promising results in last year submission and this year this idea aims to be expanded. Therefore, when possible (stereo audio is provided as Subtask B) HPD audio representation is used being D the difference between channels $(L - R)$. This forms an audio representation of $F \times T \times 3$ shape being F the number of frequency bins and T the number of temporal bins. For Subtask A, as audio is provided in mono, HP is used as log-Mel audio representation with $F \times T \times 2$ shape.

Other audio representation that showed promising results last year was LRD (left-right-difference) [9]. In this submission we change the bank of filters and instead of using Mel filters, Gammatone filter bank (GT) is used [12, 13]. As this representation cannot be obtained in Subtask A, Gammatone representation of the mono signal is obtained. Finally, for Subtask B, the idea of concatenating both HPD and LRD is explored creating a 6 channel audio representation.

All representations are calculated with a window size of 40 ms with 50% overlapping. For Subtask A, representations of 64 frequency bins are obtained to analyze the network's behaviour in problems of this nature. According to past editions and state-of-the-art research, the decision of the spectral resolution can be a decisive factor [14]. For Subtask B, representations of also 64 frequency bins are calculated, as this representation show good results in last year submission in ASC with data from the same device. GT representation where implemented using the Auditory Toolbox presented in [15] with Python implementation and Mel representation using Librosa Python module [16]

2.2. Residual Squeeze-Excitation Networks

Most Acoustic Scene Classification framework rely on the ability of a CNN to extract meaningful features. Either in a VGG-style [6] or Residual [17] networks are very similar between different submissions. Therefore, the improvement of the systems often falls on other aspects such as data augmentation techniques (mixup [18] and temporal cropping among others) or the ensemble of many independent models [19, 20, 21].

In [22] an analysis of different residual-excitation blocks proposed in [23] plus the contribution of two novel blocks using the *Concurrent Spatial and Channel Squeeze and Channel Excitation* configuration presented in [24] is carried out in ASC task. The analysis is run without any data augmentation technique. According to [22], the novel *Conv-StandardPOST* configuration shows the best results in ASC problems. All networks used for this submission incorporate this residual-excitation block. For more insight of this choice, see [22].

2.3. Experimental details

All models have been trained with same configuration. The optimizer used was Adam [25] with default parameters. The models were trained with a maximum of 500 epochs. Batch size was set to 32. The learning rate started with a maximum value of 0.001 decreasing with a factor of 0.5 in case of no improvement in validation accuracy after 20 epochs. The training is considered as early finished in case of no improvement in validation accuracy after 50 epochs. Due to the competition context, mixup [18] with $\alpha = 0.4$ has been implemented. Keras with Tensorflow backend was used to implement the models of this submission.

3. SUBTASK A

In this task, it is tried to see how residual squeeze-excitation blocks can improve scene classification in the context of mismatch devices. So that to be accomplished, network presented in [22] has been used. The block used for this task has been *Conv-StandardPOST* that has shown better results than the other presented in the cited work. The number of filters, and other hyperparameters such as dropout rates, pooling sizes, activations remain the same as [22].

Input used	**Submission name**
M	Naranjo-Alcazar_Vfy_task1a_1
HP	Naranjo-Alcazar_Vfy_task1a_2

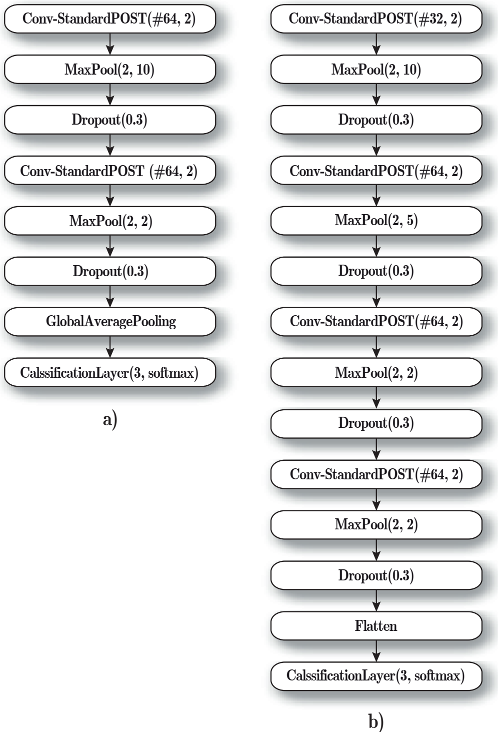
Table 1: Relationship between the systems presented in this paper and the name of the submsission for Subtask A

Model	Input	Accuracy	Log loss
Baseline		54.1	1.356
Proposed	M	**65.12**	**1.120**
	HP	61.72	1.261

Table 2: Results in development stage of the studied network and inputs for Subtask A. Accuracy and Log loss are presented as the average among classes.

As the audio of this subtask is provided in mono, the representation using the Gammatone filter bank will be done on the mono signal itself (M) and the source separation will not incorporate the difference as the third channel (HP) as explained in Section 2.

The results obtained in this subtask are shown in Table 2. The Gammatone representation, even though it only has one input channel, shows better performance than source separation. Table 1 shows the name of the submission for each system for clarity by the time the results of the challenge are published for Subtask A.


Conv-StandardPOST(#64, 2)
MaxPool(2, 10)
Dropout(0.3)
Conv-StandardPOST (#64, 2)
MaxPool(2, 2)
Dropout(0.3)
GlobalAveragePooling
CalssificationLayer(3, softmax)
Conv-StandardPOST(#32, 2)
MaxPool(2, 10)
Dropout(0.3)
Conv-StandardPOST(#64, 2)
MaxPool(2, 5)
Dropout(0.3)
Conv-StandardPOST(#64, 2)
MaxPool(2, 2)
Dropout(0.3)
Conv-StandardPOST(#64, 2)
MaxPool(2, 2)
Dropout(0.3)
Flatten
CalssificationLayer(3, softmax)

a)

b)

Figure 2: Networks submitted for Subtask B. A) network has *float32* precision while B) needs to be changed to *float16* precision to satisfy task requirements

4. SUBTASK B

4.1. Global Performance

As this subtask's aim is to propose Low-Complexity ASC framework it has been decided to analyze the depth contribution of the network by analyzing a slim *float32* precision network and a deeper *float16* precision network. Architectures can be found in Figure 2.

As it can be appreciated in Figure 2. The networks designed in this task are inspired by two previous works. The choice of various hyperparameters of the network (dropouts rates or max pooling sizes) is based on the analysis carried out last year for the same task [9] . On the other hand, the choice of the convolutional block is based on the work done in [22].

The results obtained for both networks: a) and b) are very similar, as shown in Table 3. Regarding the representation of the audio, the Gammatone representation improves substantially on the representation of the harmonic and percussive sources. The choice to change the filter bank of the LRD representation from Mel-spectrogram to Gammatone has led to an improvement of this representation (LRD). Last year submission using the Mel scale in all representations, LRD performed worse than HPD [9]. The system with LRD input exceeds the proposed baseline by 10 percentage points. By analyzing each class, the indoor, outdoor and transportation classes are improved by 12, 10 and 6 percentage points respectively.

The concatenation of both representations of 3 channels does not suppose an improvement in the classification since it obtains the similar results as LRD. As it can be observed, this representation manages to improve the result in the classification of the indoor and outdoor classes but it worsens transportation. Thus, the average is slightly lower than the LRD representation. The use of a 6-channel input forces the choice of ratio equal to 4 in the second *Conv-StandardPOST* in network a) to meet the complexity requirements. Table 5 shows the name of the submission for each system for clarity by the time the results of the challenge are published for Subtask B.

4.2. Model Complexity

As explained in Section 1 this subtask has an extra restriction as the size/complexity of the model. The limit is 500 KB for non-zero parameters. This limit disables any transfer learning technique because the networks that would be used for feature extraction must also be taken into account. In the context of this work, it has been decided to design a network that meets such requirements with float32 precision (see Figure 2a)) and a deeper one with a greater number of non-zero parameters but which has float16 precision (see Figure 2b)). For complexity calculation purposes, a script is provided by the organizers that has been used. Its outcomes are shown in Table 4.

As it can be noticed in the Table 4, all the models meet the complexity requirements. Regarding the *float32* precision models, it must be taken into account that when the input has 6 channels, the second ratio has been modified to 4. Therefore, the model size is smaller. On the other hand, it is true that with precision *float16*, there are still some free KB. However, it has been observed that deeper networks than this show worse results, being very prone to overfitting.

Model	Input	Accuracy				Log-Loss			
		indoor	outdoor	transportation	avg	indoor	outdoor	transportation	avg
Baseline		82.0	88.5	91.5	87.3	0.680	0.365	0.282	0.437
a)	HPD	92.2	90.5	97.0	93.2	0.258	0.344	0.191	0.270
	LRD	94.9	98.6	**97.7**	**97.1**	0.189	0.105	**0.107**	0.132
	LRDHPD	**95.7**	**98.8**	96.6	97.0	**0.131**	**0.063**	0.129	**0.104**
b)	HPD	90.7	95.6	95.7	94.0	0.362	0.239	0.258	0.283
	LRD	95.4	98.0	97.4	96.9	0.161	0.107	0.140	0.134
	LRDHPD	94.1	97.9	97.5	96.5	0.220	0.132	0.164	0.172

Table 3: Results in development stage of studied networks and inputs for Subtask B

Model	N^{o} Channels	Precission	Total Size
Baseline	1	*float32*	450 KB
A	3	*float32*	496.3 KB
	6		495.8 KB
B	3	*float16*	477.2 KB
	6		479.1 KB

Table 4: Complexity of the models submitted to Subtask B

Input used	**CNN**	**Submission name**
LRD	a)	Naranjo-Alcazar_Vfy_task1b_1
HPDLRD	a)	Naranjo-Alcazar_Vfy_task1b_2

Table 5: Relationship between the systems presented in this paper and the name of the submsission for Subtask B

5. CONCLUSION

For Subtask A, the Gammatone representation has been shown to perform better than source separation. The presented network improves the baseline by 11 percentage points. However, the mismatch devices problem requires a more meticulous study when it comes to representing the audio or when dealing with such representation internally over the network. It is hoped that a more optimal representation in this case plus the blocks presented in this work may show better results in the future.

Regarding Subtask B, it has been demonstrated that the representations studied in the previous submission (HPD and LRD) are suitable for the low-complexity ASC task. The LRD representation has been modified with the Gamamtone filter bank. As far as the network is concerned, it has been proven that extremely deep networks are not necessary for this task and with the datasets available. It can be concluded that when there are some limitations regarding the neural network, the point of study is the representation of the audio. The same network has shown different results, in fact, 4 percentage points of difference according to the representation used (HPD vs LRD in network a)). Unlike the image field, in the audio field, CNNs need a 2D representation that enhances the events to be detected or classified. This step is not trivial and should not be overlooked by solutions that obviate this problem by making huge ensembles or generating infinite samples to train the network. One of the challenges for the ASC right now is the implementation of these systems in a real-time environment taking into account existing limitations. In this case, audio representation plays a crucial role in achieving the shallowest possible networks.

6. 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.

7. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, “Acoustic scene classification: an overview of dcase 2017 challenge entries,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 411–415.

[2] ——, “Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups,” 2019.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, “Dcase 2016 acoustic scene classification using convolutional neural networks,” in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[4] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, “Acoustic scene classification: Classifying environments from the sounds they produce,” *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[5] A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 9–13. [Online]. Available: https://arxiv.org/abs/1807.09840

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

[7] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[8] Y. Aytar, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” in *Advances in neural information processing systems*, 2016, pp. 892–900.

[9] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, M. Cobos, and F. J. Ferri, “Cnn depth analysis with different channel inputs for acoustic scene classification,” *arXiv preprint arXiv:1906.04591*, 2019.

[10] D. Fitzgerald, “Harmonic/percussive separation using median filtering,” in *Proc. of DAFX*, vol. 10, no. 4, 2010.

[11] J. Driedger, M. Müller, and S. Disch, “Extending harmonic-percussive separation of audio signals.” in *ISMIR*, 2014, pp. 611–616.

[12] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, “Investigating the use of a gammatone filterbank for a cochlear implant coding strategy,” *Journal of neuroscience methods*, vol. 277, pp. 63–74, 2017.

[13] Z. Zhang, S. Xu, S. Cao, and S. Zhang, “Deep convolutional neural network with mixup for environmental sound classification,” in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[14] K. J. Piczak, “The details that matter: Frequency resolution of spectrograms in acoustic scene classification,” *Detection and Classification of Acoustic Scenes and Events*, pp. 103–107, 2017.

[15] M. Slaney, “Auditory toolbox,” *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.

[16] 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 *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[17] 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.

[18] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.

[19] Y. Han, J. Park, and K. Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” *the Detection and Classification of Acoustic Scenes and Events (DCASE)*, pp. 1–5, 2017.

[20] H. Chen, Z. Liu, Z. Liu, P. Zhang, and Y. Yan, “Integrating the data augmentation scheme with various classifiers for acoustic scene modeling,” *arXiv preprint arXiv:1907.06639*, 2019.

[21] K. Koutini, H. Eghbal-zadeh, and G. Widmer, “Cp-jku submissions to dcase’19: Acoustic scene classification and audio tagging with receptive-field-regularized cnns,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[22] 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.

[23] 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

[24] 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.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.