THE UNIVPM-INRIA SYSTEMS FOR THE DCASE 2020 TASK 4

Technical Report

Samuele Cornell[1*], *Giovanni Pepe*[1*], *Emanuele Principi*[1*], *Manuel Pariente*[2*], *Michel Olvera*[2*], *Leonardo Gabrielli*[1], *Stefano Squartini*[1]

[1] Università Politecnica delle Marche, Dept. Information Engineering, Ancona, Italy, {s.cornell;g.pepe}@pm.univpm.it, {e.principi;l.gabrielli;s.squartini}@univpm.it
[2] INRIA Nancy Grand-Est, Dept. Information and Communication Sciences and Technologies, France {manuel.pariente;michel.olvera}@inria.fr

ABSTRACT

In this technical report, we propose different Sound Event Detection (SED) systems for the 2020 DCASE Task 4 challenge. Given the mismatch between synthetic labelled data and target domain data, we exploit a domain adversarial training to improve the network invariance to different types of background noise. Furthermore, we use dynamic mixing and augmentation of synthetic examples at training time as well as prediction smoothing by using Hidden Markov Models. In one system, we also show that using a learnable dynamic compression, Per-Channel Energy Normalization (PCEN) front-end improves robustness to background noise by making it Gaussian. Finally, an ensemble of models proves beneficial to improve the prediction score. Concerning joint separation and sound event detection we propose a permutation-invariant training scheme to optimize directly the Sound-Event-Detection objective.

Index Terms— Sound Event Detection, Domain Adversarial Training, Per-Channel Energy Normalization, Source Separation, End-to-End

1. INTRODUCTION

The DCASE 2020 Task 4 challenge offers the opportunity to tackle Sound Event Detection (SED) in domestic environments facing real-world issues such as weakly-annotated data, unlabeled data and only a very small corpus of strongly annotated, synthetic data. The datasets are unbalanced and diverse. Specifically, the DESED dataset offers, for training, real soundscapes with weak or no labels, and isolated synthetic events with strong labels. The SINS and TUT Acoustic scenes 2017 datasets offer background noise. The source separation dataset offers isolated events but no annotations. Indeed source separation is one of the novel challenges posed in Task 4 for 2020: proposed algorithms can exploit source separation to test whether this can improve SED.

In this work we propose our strategy for SED in the context of the DCASE 2020 Task 4 challenge. Our strategy aims at pushing previous results further by exploiting data manipulation, pre-processing and post-processing techniques and adversarial techniques in the framework of the well established mean-teacher CNN+RNN (CRNN) network obtaining the first position in DCASE 2018 [1] and established, with some variations, as a baseline for Task 4 in 2019 and 2020 [2].

[*] Equal contributions

2. SOUND-EVENT DETECTION (SCENARIO 1)

In this section we will illustrate the techniques employed in our submitted systems for Sound-Event-Detection. We started from the baseline code and kept the CRNN-based architecture as well as the mean-teacher training scheme with same hyper-parameters. Our main contributions are thus in the training procedure, on the feature pre-processing and on the prediction post-processing and smoothing. Regarding training procedure, we achieved good validation set results by combining Domain Adversarial Training with online creation of synthetic labeled examples. This combined with Hidden Markov Model prediction smoothing allowed us to achieve 45.2 % event-based macro F1 score on the validation set. We also explored feature pre-processing by employing several parallel Per-Channel Energy Normalization front-end layers [3].

2.1. Domain-Adversarial Training

The datasets include data from different domains, including real and synthetic ones. This can be problematic, especially if the training and the target domains are different. Domain Adversarial Training [4] (DAT) provides a solution to this by enforcing a model to learn features that are invariant to the change of domains. This is achieved by embedding the domain adaptation process into the training procedure by adding, to the original architecture, a branch with a gradient reversal layer followed by a domain classifier. The added branch is only used at training time and then dropped at test-time, so there is no computational overhead at run-time. During training, both the network and the added domain classifier are jointly optimized. The gradient reversal layer encourages the feature extraction stage of the original architecture to work adversarially to the added domain classifier by extracting features that are domain-invariant and thus maximize the loss of the domain classification task.

Actually in our implementation we did not use the gradient reversal layer proposed by [4], but a two step optimizing procedure akin to the one used in Generative Adversarial Networks [5].

We employed a modified version of Conv-TasNet [6] separator network for the adversarial branch. We used the implementation available from Asteroid source separation toolkit [7]. In our modified version the separator network, instead of outputting a mask for each transformed-domain feature bin, it outputs a probability on the whole input example by using mean pooling. In fact, the network must classify whether the input example belongs to synthetic examples or to weak/unlabeled examples. The adversarial branch

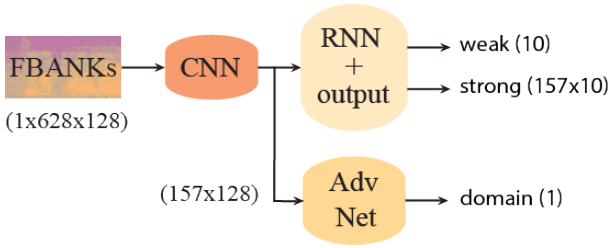

Figure 1: Domain adversarial training scheme for the CRNN architecture.

was placed in parallel to the RNN block after the CNN layers in the CRNN architecture. The whole scheme is illustrated in Figure 1.

The two networks, CRNN and adversarial, are then updated in two different steps adversarially. We denote with $\mathcal{L}_{main}$ the loss for the CRNN training, comprised of strongly labeled loss, weak labeled loss and consistency loss between teacher and student. Thus for the CRNN the update rule for its parameters θ_c is:

$$\theta_c \leftarrow \theta_c - \alpha \left(\lambda \frac{\partial \mathcal{L}_{main}}{\partial \theta_c} - (1-\lambda) \frac{\partial \mathcal{L}_{adv}}{\partial \theta_c} \right), \quad (1)$$

where $\mathcal{L}_{adv}$ is the binary cross-entropy loss for the adversarial network, λ is an hyper-parameter which controls the relative magnitude of the two losses and α is the learning rate. Differently, for the adversarial network with parameters θ_a the update rule is:

$$\theta_a \leftarrow \theta_a - l_r(1-\lambda) \frac{\partial \mathcal{L}_{adv}}{\partial \theta_a}. \quad (2)$$

We found this two-step approach to give better results than the gradient reversal layer approach, as it leads to more stable gradients during training. We tuned λ on the validation set and found that a value of 0.1 gave the best results.

2.2. Dynamic Mixing and Augmentation

Because of limited amount of acoustic diversity in DESED synthetic examples we also employed an online augmentation strategy. Each synthetic training example is constructed at training time by randomly sampling from one to five random foregrounds and one background file from SINS. We apply reverberation to each source independently by using FUSS Room Impulse Responses (RIRs). Then we apply a random time-domain augmentation chain with different effects to each source, with a maximum of two random cascaded effects:

- additive noise bursts;
- additive sine bursts;
- time-varying comb filters;
- compression;
- pitch shifting;
- low-pass and high-pass filtering.

Finally we mix the foregrounds and background. The level for each foreground is randomly sampled between -35 dB and 0 dB while the background is constrained to be at max 5 dB over the foreground with minimum level. On the feature domain we add gaussian noise with SNR between -30 dB and 10 dB and we employ SpecAugment [8]. This procedure ensures a virtually infinite amount of different strongly labelled data.

For weak and unlabeled data, we use a slightly different augmentation scheme as the foregrounds and backgrounds are not available. We only randomly add an additional background from SINS to the original mixture in time domain with 50% probability and employ only the aforementioned feature domain augmentations. In fact, we found that using time-domain augmentations on this data actually worsened the performance as the network failed to generalize to validation set when weak and unlabeled data was strongly augmented.

2.3. Per-Channel Energy Normalization

We also experimented with Per-Channel Energy Normalization (PCEN) as a learnable dynamic compression strategy. This technique is able to enhance transient audio events while transforming many soundscape noise patterns into additive white gaussian noise improving the robustness of audio classification algorithms in presence of background noise [9]. While this operation can be helpful to enhance some sound events in domestic environments, the filtering operation involved in the computation of PCEN can have a negative impact in sound classes with slowly varying spectro-temporal characteristics, for instance, vacuum cleaner or blender events. Therefore, instead of learning the parameters of a single transformation that finds a trade-off between standing out fast transition sounds and not degrading the quality of stationary-like sounds, we propose to learn several PCEN transformations in parallel. We feed the outputs as feature channels to the CRNN model and jointly optimize the parameters of such PCEN layers using backpropagation. While PCEN was originally proposed on mel-energies, we found here, that, using log-mel energies resulted in slightly better validation performance than using mels. We found also that a number of layers of two was sufficient to give significant performance improvement and that adding more layers did not bring additional appreciable improvement.

2.4. HMM smoothing

Hidden-Markov-Model (HMM) decoding was employed to obtain final predictions instead of the simple median filtering scheme used in the baseline. A two state HMM was employed for each class. The silence self-loop transition probability was tied to be the same for all HMMs. We tuned the self-loop transition probabilities for every class and silence on the validation set using a 50% split by using Random Forest and with the objective of maximizing the event-based F1 macro-average score of the trained CRNN model. Once found the optimal parameters for the HMMs transition probabilities, inference is performed by running Viterbi decoding on the CRNN-obtained emission probabilities for each class.

2.5. SED-Results

Hereafter we report our results obtained on validation set for our submitted systems. Each submitted system is comprised of combinations of aforementioned techniques and will be described shortly. Unless stated otherwise we used identical hyper-parameters and techniques as in the baseline system. We submitted two single systems:

- **PCEN** this system comprises of the baseline system with on top two parallel PCEN layers applied directly on log-mel features. For this system we did not use any data-augmentation strategy. In fact, we found that data-augmentation hampered performance when used in conjunction with PCEN layers. We also used a global min-max normalization scheme instead of the mean and variance normalization scheme used in the baseline. HMM prediction smoothing is performed on output probabilities.
- **DAT+HMM** this system comprises of baseline system plus Domain Adversarial Training, HMM prediction smoothing and online mixing and augmentation. We use the same normalization scheme as it is used in baseline system by computing mean and variance of non-augmented features over all training data.

Finally, we also submitted two ensemble systems: **Ensemble DAT+PCEN** and **Ensemble DAT+PCEN HMM 2**. For these two submissions we actually used the same models. The only difference is in different HMM transition probabilities. We used an ensemble of three different single systems: PCEN and two DAT models from two different training runs. To obtain emission probabilities we simply averaged the outputs of the different models.

Table 1: Performance of submitted SED systems on validation set.

Method	Event macro F1 score	PSDS
Baseline	34.8	0.61
PCEN[1]	43.69	0.63
DAT+HMM	45.2	0.68
Ensemble DAT+PCEN	46.17	0.69
Ensemble DAT+PCEN HMM 2	47.44	0.69

2.6. Ablation Study

In Table 2 we compare results obtained on validation set by the challenge baseline system and results obtained by adding the proposed techniques to the challenge baseline. We can see that PCEN alone is able to bring substantial improvement in performance. Instead, online mixing and augmentation (Augm) brings modest performance improvement on its own. We suspect this is due to the fact that the online generated examples are only partially representative of true target-domain sound events and thus the network still can incur in overfitting of synthetic examples. It however brings significant benefits when it is coupled with DAT. On the other hand, HMM smoothing alone is able to constantly give at least two points performance improvement on all systems, with the improvement being greater as the model predictions get more reliable.

3. SED-AWARE SEPARATION (SCENARIO 3)

We also experimented with Source-Separation as a pre-processing step for SED. We did not perform joint End-to-End training of SED and separation, instead, we trained the separation model using the pre-trained SED baseline. We thus performed End-to-End training in order to optimize the separation model directly for the SED task, but, we did not update the SED model whose weights were kept frozen. The proposed approach is significantly different from the one employed by the challenge official source separation baseline. The official source separation baseline is derived from [10] and is trained on a synthetic dataset comprised of FUSS foregrounds and SINS background as well as foregrounds from DESED. This baseline model is optimized with a Scale-Invariant Signal-to-Noise-Ratio (SI-SNR) [11] to remove the background noise from the mixtures, thus performing denoising rather than full foregrounds separation from the mixtures.

[1]submission label for this system is Ensemble MT+PCEN but it is actually only baseline CRNN system plus PCEN front-end and HMM post-processing

Table 2: Ablation study for proposed techniques.

Method	Event macro F1 score
Baseline	34.8
+HMM	37.13
+PCEN	39.93
+PCEN+HMM	43.69
+Augm	37.31
+DAT+Augm	40.91
+DAT+Augm+HMM	45.2

This approach does not guarantee that the denoised mixtures will be more suitable for SED. In fact, the denoising process could lead to mixtures whose distribution is significantly different from the one of noisy mixtures, which are used to train the SED model. Thus the denoising process can potentially introduce a mismatch. This can explain the modest performance improvement given by the baseline separation model.

Our approach is illustrated in Figure 2. We perform Deep Neural Network (DNN) mask-based separation directly on mel-spectrograms. The separated features are then fed to the SED baseline after applying logarithm and scaling. We then use both the predictions of the SED as well as its internal activations to train the mask-estimation DNN network. We use the same data used for baseline SED training, comprised of synthetic (for which foregrounds are available), weakly and unlabelled examples. For synthetic examples we used dynamic mixing of sources as explained in Section 2.2 However we did not employ any augmentation strategy apart from this. In fact we found that applying time-domain and feature-domain augmentation as in Section 2.2 led to worse performance. Permutation Invariant Training [12, 13] and Mean Teacher[14] are used to train the mask-estimation DNN with different losses:

- **Strongly Labelled**: for synthetic data, for which foregrounds are available we compute permutation-invariant Mean-Squared Error loss $\mathcal{L}_{MSE}$ and find the optimal permutation for the estimated separated mixtures. The reordered estimated foregrounds features are then fed to the SED model and we compute deep feature loss [15, 16] $\mathcal{L}_{dfl}$ between activations obtained with estimated foregrounds and those obtained with oracle foregrounds.
- **Weakly Labelled**: for weakly labelled data no oracle foregrounds are available, thus we train the separation model to minimize permutation-invariant binary cross entropy between weak predictions of SED model when it is fed the estimated foregrounds features and the weak labels.
- **Mean-teacher consistency**: we use the Mean Teacher method for the mask-estimation network and enforce SED weak and strong predictions consistency between the values obtained with student separation model and an exponential moving aver-

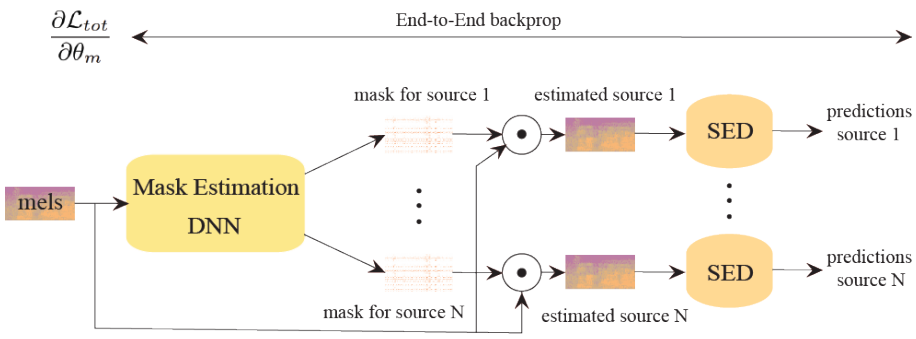

Figure 2: End-to-End training for SED-aware Separation.

age mean teacher separation model using permutation invariant MSE loss $\mathcal{L}_{teach}$.

For mask-estimation we used a reduced version of the separator network from Conv-TasNet [6], as implemented in Asteroid, with 5 blocks ($X = 5$) and 3 repeats ($R = 3$), 64 bottleneck channels, 128 depth-wise convolution channels and sigmoid mask. The whole system was trained to separate a maximum of 5 different sound event classes. We use a batch size of 32 examples with respectively 12 synthetic examples, 12 weakly labelled examples and 8 unlabelled examples.

3.1. Results

In Table 3 we report results obtained with submitted separation system **separation_hmm** on validation data in terms on event-based F1 score. For prediction smoothing we used HMM smoothing described in Section 2.4 instead of median filter. We also report here F1 score without HMM smoothing and with same median filter smoothing used in the baseline. It can be seen that even without HMM smoothing the proposed system significantly outperforms the baseline.

Table 3: Performance of submitted separation system on validation set.

Method	Event macro F1 score
Baseline	35.6
separation_hmm	40.16
separation	37.02

4. CONCLUSIONS

In this work we outlined our proposed techniques for tackling the 2020 DCASE Task 4 challenge. Instead of focusing on neural architecture we experimented directly with the baseline. Regarding SED, we showed that, at least on validation set, the addition of PCEN front-end feature pre-processing, Domain Adversarial Training and online data augmentation and mixing can bring substantial benefits with null or insignificant computational overhead at inference time. As another contribution we also showed that HMM smoothing alone can greatly improve performance of the systems by refining network predictions.

Regarding separation we experimented with pre-trained SED system and we did not performed joint separation and SED system training. Nevertheless we opted for a End-End approach where we used the frozen SED model predictions to drive the separation network training. Separation was thus performed directly in feature domain by a mask-based approach. In this way the mask-estimation separation DNN learns directly to separate the input mixture in a way that minimizes the SED objective (strong and weak). This approach shows notable improvement over the source separation baseline which instead is trained with a source-separation objective.

5. ACKNOWLEDGMENTS

Some experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

6. REFERENCES

[1] L. JiaiKai, "Mean teacher convolution system for dcase 2018 task 4," in *DCASE 2018 Tech Report*.

[2] L. Delphin-Poulat and C. Plapous, "Mean teacher with data agumentation for dcase 2019 task 4," in *DCASE 2019 Tech Report*.

[3] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.

[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, *et al.*, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[8] 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:1904.08779*, 2019.

[9] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.

[10] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *in preparation*, 2020.

[11] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[12] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[13] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[14] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[15] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[16] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.