ONSETS, ACTIVITY, AND EVENTS: A MULTI-TASK APPROACH FOR POLYPHONIC SOUND EVENT MODELLING

Arjun Pankajakshan[1], *Helen L. Bear*[1], *Emmanouil Benetos*[1,2],

[1] School of EECS, Queen Mary University of London, UK [2] The Alan Turing Institute, UK
{a.pankajakshan, h.bear, emmanouil.benetos}@qmul.ac.uk

ABSTRACT

State of the art polyphonic sound event detection (SED) systems function as frame-level multi-label classification models. In the context of dynamic polyphony levels at each frame, sound events interfere with each other which degrade a classifier's ability to learn the exact frequency profile of individual sound events. Frame-level localized classifiers also fail to explicitly model the long-term temporal structure of sound events. Consequently, the event-wise detection performance is less than the segment-wise detection. We define 'temporally precise polyphonic sound event detection' as the subtask of detecting sound event instances with the correct onset. Here, we investigate the effectiveness of sound activity detection (SAD) and onset detection as auxiliary tasks to improve temporal precision in polyphonic SED using multi-task learning. SAD helps to differentiate event activity frames from noisy and silence frames and helps to avoid missed detections at each frame. Onset predictions ensure the start of each event which in turn are used to condition predictions of both SAD and SED. Our experiments on the URBAN-SED dataset show that by conditioning SED with onset detection and SAD, there is over a three-fold relative improvement in event-based F-score.

Index Terms— Polyphonic sound event detection, sound activity detection, onset detection, multi-task learning.

1. INTRODUCTION

Sound event detection (SED) [1] is the task of detecting the label, onset, and offset of sound events in audio streams. State of the art convolutional recurrent neural network (CRNN) based polyphonic SED systems use a frame-wise cost function for training [2, 3, 4]. The frame-level classifier performance depends on the dynamic polyphony level, masking effects between the sound events and the amount of co-occurrence of sound events in the training data. Frame-level classifiers also fail to explicitly model the long-term temporal structure of sound events. Due to these limitations, frame-level training methods are not sufficient to model the overall acoustic features of polyphonic sound events. Consequently, the event-based detection performance is very poor compared with the segment-based detection of sound events. For example, in the DCASE 2016 task on event detection in real life audio [5], the F-score is around 30% at segment level (frame length of 1 second), but only around 5% at event level (tolerance of 200 ms for onset and 200 ms or half length for offset). Here, we define 'temporally precise polyphonic sound event detection' as the subtask of detecting sound event instances with the correct onset. In applications like audio surveillance [6] and health care monitoring [7], temporally accurate event-based detection is very important.

1.1. Related work

SED is related to the speech processing tasks of automatic speech recognition and speaker diarization, as well as the music signal-related task of automatic music transcription [8, 9]. Many sequence modelling methods in speech and music have been utilized in environmental sound event modeling. For example, in [10] hidden semi-Markov models separately model the duration of sound events; Wang and Metze used a connectionist temporal classification (CTC) cost function in a sequence-to-sequence model for SED [11]. Unlike speech and music language modelling there is not a well defined structure for environmental sound events. Explicit use of sequential information to improve sound event modelling is investigated in [12]; the co-occurrence probabilities of different events are modelled using N-grams which in turn smooth the spiky output of a neural net based SED system trained using CTC loss. However the SED performance is not much improved, considering the addition of both N-grams and the CTC loss. In [13] a hybrid approach that combined an acoustic-driven event boundary detection for sound event modelling with a supervised label prediction model is proposed for SED. This hybrid approach significantly improved event-based detection accuracy. It is assumed that the method requires additional post-processing to combine the event boundary information with the label predictions since both models are trained independently.

1.2. Contributions of this work

Within the limits of frame-wise training approaches using CRNN models for polyphonic SED, we propose a novel sequence modelling method using onset detection and SAD as auxiliary tasks to achieve temporally precise polyphonic SED using multi-task learning. SAD is the task of detecting the presence or absence of any sound events, which is analogous to voice activity detection in speech processing. The effectiveness of SAD to improve polyphonic SED is discussed in a preliminary work by the authors [14]. Unlike polyphonic SED, SAD is not affected by masking effects and acoustic variations in the sound events even when it is trained using frame-wise cost functions. The onset detector helps to predict the beginning of sound events accurately, thus reducing missed event detections which improves temporally precise SED. Both onset detection and SAD are not overwhelmed by polyphonic structure, instead these auxiliary tasks can exploit the polyphonic nature to improve temporal precision in SED. Inspired from the success of

AP is supported by a QMUL Principal's studentship. EB is supported by RAEng Research Fellowship RF/128 and a Turing Fellowship. This research was supported by an NVIDIA GPU Grant.

https://doi.org/10.33682/sm6r-8p49

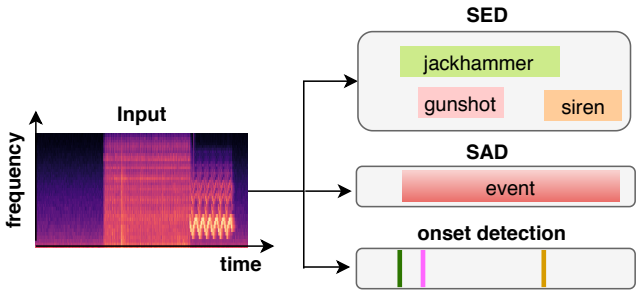

Figure 1: Block diagram of SED, SAD, and onset detection.

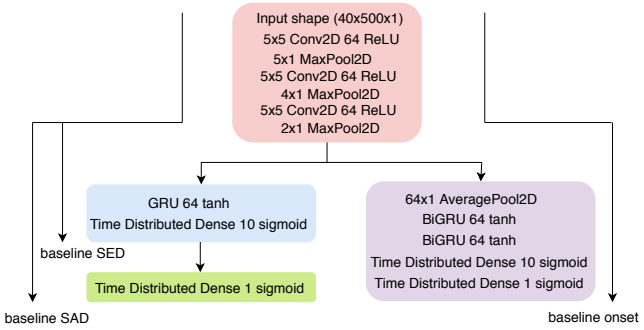

Figure 2: Baseline architectures for onset detection, SAD and SED.

framewise note detection in piano transcription conditioned on onset predictions [15], we configure multi-task models for SED in a similar fashion. We investigate the individual effectiveness of conditioning onset detection and SAD on SED. Also we propose a joint model to improve the temporal precision of sound events in polyphonic SED.

2. PROPOSED METHOD

In this work, we investigate the effectiveness of onset detection and SAD to improve the temporal precision of polyphonic SED. Onset detection exclusively predicts the beginning of a sound event instance, which is useful because many sound event onsets are characterized by sudden increase in energy, e.g. percussive sound events. The onset predictions are used to condition framewise SED in a similar way as the music transcription in [15]. Conditioning SED based on onset predictions helps to precisely locate the beginning of sound events. SAD predicts whether any event activity is present or not in each frame of the audio and so avoids the pitfalls caused by masking effects between co-occurring sound events. Furthermore, SAD can exploit polyphony to ensure the presence of an event even if one event is masked by the occurrence of another event with similar or different acoustic properties. Hence conditioning SED with SAD helps to avoid missed detections in complex acoustic conditions such as real-world sound scenes. Fig. 1 shows the complete system.

We use a state-of-the-art CRNN model architecture ([2]) to build baseline SED, SAD, and onset models. To evaluate the individual effect of each auxiliary task on SED, we implement and analyse separate SED models conditioned on onset prediction and sound activity prediction using a multi-task learning set up.

- sed_onset is the SED model conditioned on onset prediction.
- sed_sad is the SED model conditioned on sound activity.
- The sad_onset model verifies the effect of SAD conditioned on onset detection.
- The joint SED model using both onset detection and SAD as auxiliary tasks, is denoted as sed_sad_onset.

2.1. Motivation for onset detection

We examine the importance of onset detector as an auxiliary task in polyphonic SED in different perspectives. Firstly, the onset detector is able to predict the beginning of sound events more effectively compared to a standalone baseline SED model which is affected by dynamic polyphony levels and acoustic variations of the sound events. Secondly, consider the case of two closely occurred sound events. Even if the onset detector could only detect either of the events, the same prediction can be used as two separate onsets while conditioning the event detection with minimum error. This way the conditioned SED model can avoid missed detection of sound events and ensure the detection of two closely occurred events even if the onset detector actually predicted only a single onset. We evaluate the onset models based on this assumption. More precisely, there is a one-to-many relation between the onset prediction and the reference. We consider this fact when counting the false negatives for onset model evaluation and verified this relaxation does not make much difference.

2.2. Model configuration

The detailed network architecture of the three baseline models is in Fig. 2. We replicate the SED and SAD implementation from [14] and extend it with our onset detector. The onset detector has three blocks of convolutional layers. The output activations from the third convolutional layer are averaged using an average pooling layer, followed by two bidirectional Gated Recurrent Unit (GRU) layers and a fully connected sigmoid to output the onset predictions. Onsets are very localized sound events; hence the inter feature map representations learned at the final convolution layer are equally important to predict onsets so we opted for average-pooling across the third convolution layer feature maps instead of max-pooling. Bidirectional recurrent layers are proven to work well for musical onset detection [16]. The output of the SED model is a posteriogram matrix with dimensions $T \times C$, where T is the number of frames in the input data representation and C is the total number of sound event classes in the dataset. The output representation of the sound activity detector and the onset detector is a posteriogram vector with dimension T. The baseline model predictions are binarised with a threshold before evaluation. We investigate different threshold values on the baseline models using the validation set. Using the best results, we chose a threshold of 0.2 for the SED and onset predictions and 0.5 for the SAD predictions.

We implement conditional models for SED (sed_sad, sed_onset, sed_sad_onset) and SAD (sad_onset) using the baseline SED, SAD, and onset models in a multi-task joint training setup. Motivated from the effectiveness of piano note transcription conditioned on onsets, initially we implement our conditional models as explained in [15]. However, in our case the exact same architecture was not effective so we explore the possibilities of multi-task learning [17, 18, 19, 20]. From our experiments we fix our

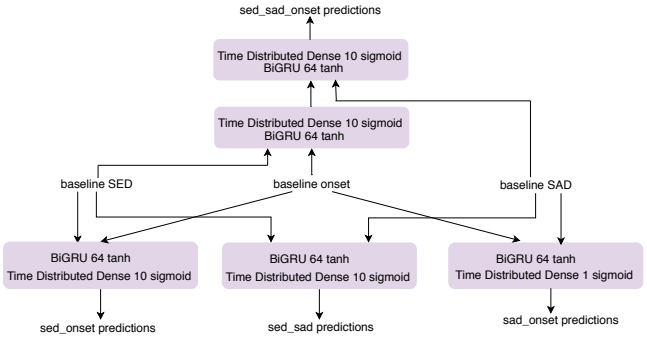

Figure 3: Block diagram of the proposed conditional models.

conditional model architectures by sharing the initial two convolutional layers of the respective tasks. The conditional model training is described in Section 2.4.

- We implement the SED model conditioned on activity detection (sed_sad) by concatenating the predictions of the SAD baseline model with the output of the baseline SED model, followed by a bidirectional GRU layer and a fully connected sigmoid layer to predict the sound events.
- The predictions of the onset baseline model are concatenated with the output of the baseline SED architecture, followed by a bidirectional GRU layer and a fully connected sigmoid layer to implement the SED model conditioned on onset detection (sed_onset). Similarly, we implement the SAD model conditioned on onset detection (sad_onset).

In the sad_onset model the activity detection is conditioned using the reference event onsets which are different from the onsets for sound activity. For example, if two sound events are overlapping with each other, there are two sound event onsets but only a single sound activity onset. The additional onsets in the event onset reference condition unwanted preference to the corresponding activity frames which cause slight disturbance of activity detection around the respective frames. As a result of this the segment-based F-score for activity detection of the sad_onset model is lower than the baseline SAD (shown in Section 4). In our analysis of the sed_onset model and the sed_sad model we realize that, conditioning SED using both onset detection and activity detection improves SED temporal precision. So we design the final joint conditioned model (sed_sad_onset) by first conditioning SED using onsets and then the onset-conditioned event prediction is reconditioned using activity predictions. Fig. 3 is a block diagram of the conditional models.

- For the sed_sad_onset model, the predictions of the onset baseline model are concatenated with the baseline SED outputs, followed by a bidirectional GRU layer and a fully connected sigmoid layer. The output from this sigmoid layer is concatenated with the predictions of the activity detector and passed through a bidirectional GRU layer and a fully connected sigmoid layer.

2.3. Feature extraction

We use librosa [21] to compute mel-scaled spectrograms from the input audio. The short-term Fourier transform (STFT) is employed to obtain the spectrogram from the input audio recordings with a hop length of 882, an FFT window of 2048, and a sample rate of 44.1 kHz. This process converts a ten second (duration of recordings in the URBAN-SED [22] dataset) audio recording into a 1024×500 dimensional spectrogram representation. Each frame of this spectrogram is converted into a 40-dimensional vector of log filter bank energies using a Mel filterbank. We apply min-max normalization on the mel band energies. Hence, each 10-second audio recording is represented by a 40×500 Mel-spectrogram.

2.4. Training

We train all the models in a supervised manner. The dimension of the labels for the SED is $T \times C$, and for the SAD and the onset detection, it is T. The training process for the SED and the SAD models is explained in our previous work [14]. For the onset detection, a single frame is used to mark each onset during the training process. The baseline models for SED, SAD and onset detection are trained using the respective cross-entropy losses denoted by L_{sed}, L_{sad}, and L_{onset}. The total loss of each of the conditional models is the weighted sum of the two corresponding cross-entropy losses as listed in (1). During training of the conditional models, the individual losses are equally weighted with a factor of 0.5.

$$\begin{aligned} L_{sed\_sad} &= 0.5 \cdot L_{sed} + 0.5 \cdot L_{sad} \\ L_{sed\_onset} &= 0.5 \cdot L_{sed} + 0.5 \cdot L_{onset} \\ L_{sad\_onset} &= 0.5 \cdot L_{sad} + 0.5 \cdot L_{onset} \\ L_{sed\_sad\_onset} &= 0.5 \cdot L_{sed} + 0.5 \cdot L_{sad} + 0.5 \cdot L_{onset} \end{aligned} \tag{1}$$

Every CNN layer activations are batch normalised [23] and regularised with dropout [24] (probability = 0.3). We train the network for 200 epochs using a binary cross entropy loss function for both tasks and with Adam [25] optimizer with a learning rate of 0.001. Early stopping is used to reduce overfitting. The proposed joint model is implemented using Keras with TensorFlow.

3. DATASET AND METRICS

We use the URBAN-SED [22] dataset in all experiments. URBAN-SED is a dataset of 10,000 soundscapes with sound event annotations generated using Scaper [22], an open-source library for soundscape synthesis. All recordings are ten seconds, 16-bit mono and sampled at 44.1kHz. The annotations are strong, meaning for every sound event the annotations include the onset, offset, and label of the sound event. Each soundscape contains between one to nine sound events from the list (air_conditioner, car_horn, children_playing, dog_bark, drilling, engine_idling, gun_shot, jackhammer, siren and street_music) and has a background of Brownian noise. The URBAN-SED [22] dataset comes with pre-sorted train, validation and test sets that we use. Among $10,000$ soundscapes, 6000 are used for training, 2000 for validation and 2000 for testing.

We use precision, recall and F-score as metrics for onset detection. For sound event and sound activity detection we use the F-score and Error Rate (ER), with F-score as the primary metric. The evaluation metrics are computed in both segment-wise and event-wise manners using the sed_eval tool [26]. Segment-based metrics show how well the system correctly detects the temporal regions where a sound event is active; with an event-based metric, the metric shows how well the system detects event instances with correct onset and offset. For temporally precise event detection, we give more importance to the event-based metric. The evaluation scores

for activity detection and event detection are micro averaged values, computed by aggregating intermediate statistics over all test data; each instance has equal influence on the final metric value. We use a segment length of one second to compute segment metrics. The event-based metrics are calculated with respect to event instances by evaluating only onsets with a time collar of 250ms.

In the case of onset detection, an onset is considered to be correctly detected if there is a ground truth annotation within ± 250ms around the predicted position. An important factor in the evaluation is how false positives and false negatives are counted [27]. Assume that two or more onsets are predicted inside the detection window around a single reference annotation. All predictions within the detection window around the single reference onset are treated as one true positive and zero false positives. The false negatives are counted by granting a one-to-many relationship between a single prediction and multiple reference onsets within an analysis window (± 250 ms). Since our main goal is to use onset detection to condition SED we believe this evaluation approach is fair.

4. EVALUATION

Tables 1, 2, and 3 show the results of onset detection, sound activity detection and sound event detection respectively. Table 1, the baseline onset detector, has the best F-score (81.68%). When onset detection is used to condition SAD and SED, the onset F-scores are slightly lower than the baseline value. However, conditioning activity detection and event detection using onsets is really effective in temporally precise SED. This is demonstrated in Tables 2 and 3. By conditioning activity detection using onsets (sad_onset in Table 2), the event-based activity detection F-score increases from the baseline value of 43.14% to 70.31%. At the same time the segment-based F-score for the same model drops from 97.48% to 70.17%. This is due to the fact that the sound event onset labels are used to condition the activity detection which is different from the actual onsets of sound activity as explained in Sec 2. We see a similar improvement when SED is conditioned using onsets (sed_onset in Table 3). For this model the event-based F-score increases from the baseline value of 7.34% to 21.42%. The segment-based F-score for the same model is 47.76% compared with the baseline value of 35.48%. The improvement in the event-based F-scores for the sad_onset model and the sed_onset model verify the effectiveness of onset conditioned polyphonic event detection to improve temporal precision in polyphonic SED.

The sed_sad model performance (in Table 3) is compared with a joint model to enhance event detection by re-weighting the event prediction using the activity prediction from [14] (sed_sad_joint in Table 3). The sed_sad model segment-based and event-based F-score values are 43.52% and 17.40% respectively; both are improvements from the baseline and sed_sad_joint model. By analysing the sed_sad and sed_onset results, we know conditioning event detection using onsets is more effective than conditioning using sound activity. To utilize the advantage of both onsets and sound activity frames in conditioning event detection we implement the final conditional model (sed_sad_onset) as explained in Sec 2. Using this model the event-based F-score improves to 23.20%.

Further analysis of onset detection performance of the conditional models (sed_onset and sed_sad_onset) reveal that when false positive errors in onset detection are less, the sound event model is more effective. More precisely, when the precision of onset detection improved from 85.96% to 89.17% from the sed_onset model to the sed_sad_onset model the event-based F-score also improved from 21.42% and 23.20%; which implies that false positive errors in the onset detection are more influential than false negative errors in the performance of conditional event detection using onsets. This analysis again proves the effectiveness of conditional sound event detection using onsets. The class-wise evaluation of all the models are available[1].

Table 1: Onset detection results.

Case	Precision	Recall	F-score
baseline	81.16	**82.20**	**81.68**
sad_onset	**90.09**	73.28	80.82
sed_onset	85.96	74.31	79.71
sed_sad_onset	89.17	70.68	78.85

Table 2: Sound activity detection results.

	F1 (%)		Error rate	
Case	Segment	Event	Segment	Event
baseline	97.48	43.14	0.05	0.78
sed_sad_joint	**98.53**	46.23	**0.03**	0.72
sad_onset	70.17	**70.31**	0.46	**0.61**
sed_sad	98.39	45.53	**0.03**	0.73
sed_sad_onset	97.58	46.51	0.05	0.76

Table 3: Sound event detection results.

	F1 (%)		Error rate	
Case	Segment	Event	Segment	Event
baseline	35.48	7.34	1.54	3.81
sed_sad_joint	41.03	8.76	0.97	3.58
sed_sad	43.52	17.40	0.88	1.68
sed_onset	**47.76**	21.42	1.02	2.33
sed_sad_onset	44.12	**23.20**	**0.85**	**1.49**

5. CONCLUSIONS AND FUTURE WORK

Within the limits of frame-wise training in sequence modelling problems, we proposed a novel sequence modelling method for temporally precise polyphonic sound event detection conditioned on onsets and sound activity detection. From our experimental results we conclude that: 1) the performance of temporally precise event detection of the conditional models depends on the performance of onset and sound activity detection and also on the conditioning method. 2) conditioning the main task using auxiliary tasks and training in a multi-task set up is an effective method to improve SED performance. We believe an onset detector with precision and recall measures close to 100% can drastically improve temporal precision in SED performance. In the future, we plan to improve onset detection by modifying the onset loss with dedicated penalty terms for false positive and false negative onset predictions. Also our current conditional models are trained using an equally weighted sum of cross entropy losses of the individual tasks. Instead of this approach we plan to develop a task-dependent weighting scheme explicitly conditioning auxiliary tasks to derive a principled multi-task loss function as demonstrated in [28].

[1] http://c4dm.eecs.qmul.ac.uk/DCASE2019/class-wise_evaluation_supplementary_doc.pdf

6. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] 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.

[3] E. Çakir and T. Virtanen, "End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–7.

[4] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," *Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2017.

[5] "Sound event detection in real life audio, dcase 2016 challenge results," http://www.cs.tut.fi/sgn/arg/dcase2016/task-results-sound-event-detection-in-real-life-audio.

[6] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Reliable detection of audio events in highly noisy environments," *Pattern Recognition Letters*, vol. 65, pp. 22–28, 2015.

[7] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[8] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[9] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, 2016.

[10] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, K. Takeda, T. Hayashi, S. Watanabe, T. Toda, T. Hori, *et al.*, "Duration-controlled lstm for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 11, pp. 2059–2070, 2017.

[11] Y. Wang and F. Metze, "A first attempt at polyphonic sound event detection using connectionist temporal classification," in *ICASSP 2017-2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2986–2990.

[12] G. Huang, T. Heittola, and T. Virtanen, "Using sequential information in polyphonic sound event detection," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 291–295.

[13] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 36–40.

[14] A. Pankajakshan, H. L. Bear, and E. Benetos, "Polyphonic sound event and sound activity detection: A multi-task approach," *in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019.*

[15] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *International Conference of Music Information retrieval (ISMIR)*, 2018.

[16] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long-short term memory neural networks," in *Proc. 11th Intern. Soc. for Music Information Retrieval Conference (ISMIR)*, 2010, pp. 589–594.

[17] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning (ICML)*. ACM, 2008, pp. 160–167.

[18] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *in International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016.

[19] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[20] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 41–48.

[21] 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*, 2015, pp. 18–25.

[22] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *International Conference on Machine Learning (ICML)*, 2015.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[26] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[27] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods." in *Intern. Soc. for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 49–54.

[28] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7482–7491.