Voice activity detection in the wild via weakly supervised sound event detection

Yefei Chen[*], *Heinrich Dinkel*[*], *Mengyue Wu ,Kai Yu*

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai, China

{chenyefei,richman,mengyuewu, kai.yu}@sjtu.edu.cn

Abstract

Traditional supervised voice activity detection (VAD) methods work well in clean and controlled scenarios, with performance severely degrading in real-world applications. One possible bottleneck is that speech in the wild contains unpredictable noise types, hence frame-level label prediction is difficult, which is required for traditional supervised VAD training. In contrast, we propose a general-purpose VAD (GPVAD) framework, which can be easily trained from noisy data in a weakly supervised fashion, requiring only clip-level labels. We proposed two GP-VAD models, one full (GPV-F), trained on 527 Audioset sound events, and one binary (GPV-B), only distinguishing speech and noise. We evaluate the two GPV models against a CRNN based standard VAD model (VAD-C) on three different evaluation protocols (clean, synthetic noise, real data). Results show that our proposed GPV-F demonstrates competitive performance in clean and synthetic scenarios compared to traditional VAD-C. Further, in real-world evaluation, GPV-F largely outperforms VAD-C in terms of frame-level evaluation metrics as well as segment-level ones. With a much lower requirement for frame-labeled data, the naive binary clip-level GPV-B model can still achieve comparable performance to VAD-C in real-world scenarios.

Index Terms: Voice activity detection, semi-supervised learning, deep neural networks, sound event detection

1. Introduction

Voice activity detection (VAD), whose main objective is to detect voiced speech segments and distinguish those from unvoiced ones, is a crucial component for tasks such as speech recognition, speaker recognition, and speaker verification. Deep learning approaches have been successfully applied to VAD [1, 2, 3, 4]. For VAD in complex environments, neural networks (NN) have been successful. Deep neural networks (DNN) and specifically convolutional neural networks (CNN) offer improved modeling capabilities compared to traditional methods [2], while recurrent- (RNN) and long short-term memory (LSTM) networks can better model long-term dependencies between sequential inputs [1, 5, 6]. However, despite the application of deep learning methods, NN-based VAD training still requires frame labels. Thus training data utilized is usually under controlled environment with or without additional synthetic noise [7]. This inevitably prevents VAD from real-world applications, where speech in the wild is often accompanied by countless unseen noises with different features.

Therefore, this paper intends to propose a method to detect speech beyond clean and controlled noisy environment.

It should be noted that frame-level labels are quite unlikely to come with real-world recordings since manual labeling is costly, and label predictions from a Hidden Markov model need prior knowledge about the language being spoken [8]. A task to detect speech components while enabling noisy data training, is related to weakly-supervised sound event detection (WSSED), which detects and localizes different sounds, including speech via clip-level supervision. Since WSSED systems are reported [9] to be robust to noise and only require clip-level labels, this work integrates WSSED methods in scaling VAD to speech in-the-wild scenarios and relaxing its dependence on frame labeling. Specifically, we investigate two questions: 1) Are current, multi-class WSSED models comparable in performance to DNN-based VAD; 2) Is utterance-level training a viable alternative compared to frame-level? We thus introduce our framework, a general-purpose training framework for VAD (GPVAD, see Figure 1). By general purpose, we refer to two distinct aspects: First, the framework is noise-robust and capable of being deployed in wild, real-world scenarios; Secondly, the framework can be trained on unconstrained data, thus enabling learning from big webly data like noisy online-videos.

The paper is structured as follows: In Section 2, we briefly review the related work on WSSED and how it can be transferred for VAD in the wild. In Section 3, the GPVAD approach is introduced. Moreover, in Section 4 we introduce our experimental setup and provide implementation details. In Section 5 the results are presented and finally in Section 6 a conclusion is provided.

2. Weakly supervised sound event detection

Since WSSED can work well in detecting speech in a noisy environment without frame-level labeling, we borrow this idea to realize VAD in the wild. Here we present related work on sound event detection (SED), which aims to classify (audio tagging) and possibly localize multiple co-occurring sound events from a given audio clip. In this work, we mainly focus on weakly-supervised SED (WSSED), a semi-supervised task, which has only access to clip-level labels during training, yet needs to classify and localize a specific event during evaluation. This weakly-supervised fashion enables training on noisy data with lower requirements for the labeling method. Recent advances in weakly supervised sound event detection, in particular, the detection and classification of acoustic scenes and events (DCASE) challenges [10], led to large improvements for predicting accurate sound event boundaries as well as event-labels [11, 12, 13, 14, 15, 16]. In particular, recent work [9] has shown promising performance regarding short, sporadic events such as speech.
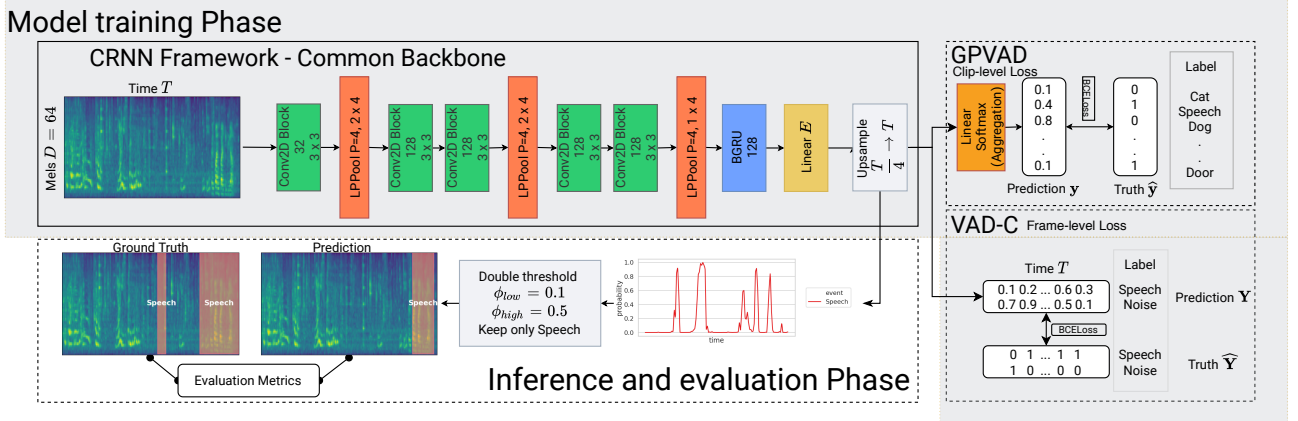
---

[*] equal contribution. Mengyue Wu and Kai Yu are the corresponding authors.

Figure 1: *The proposed framework. A CRNN architecture is utilized, while GPVAD is trained clip-level labels, and VAD-C trained on frame-level labels. Each Conv2d block represents a batch-normalization, followed by a zero-padded 2-dimensional convolution with kernel size 3×3 and a leaky ReLU activation with a negative slope of 0.1. The CNN output is fed into a bidirectional gated recurrent unit (GRU) with 128 hidden units. The architecture sub-samples the temporal dimension T by a factor of 4 and later upsampled to match the original input temporal dimension. The number of events E is set to be 527 for GPV-F, 2 for GPV-B, and VAD-C. After post-processing the output, only the Speech event is kept for final evaluation.*

3. VAD in the wild via WSSED

Traditionally, VAD for noisy scenarios is modeled as in Equation (1). The assumption is that additive noise $\mathbf{u}$ can be filtered out from an observed speech signal $\mathbf{x}$ to obtain clean speech $\mathbf{s}$.

$$\mathbf{x} = \mathbf{s} + \mathbf{u} \qquad (1)$$

However, directly modeling $\mathbf{u}$ is rather tricky, since each type of noise has its individual traits. Therefore, we aim at learning the properties of $\mathbf{s}$ by observing it with potentially L different non-speech events $(\mathbf{u}_1 \dots, \mathbf{u}_L)$. Those events are not restricted to being background/foreground noises and can have distinct real-world sounds (e.g., Cat, Music).

$$\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_l, \dots, \mathbf{x}_L\}$$
$$\mathbf{x}_l = (\mathbf{s}, \mathbf{u}_l) \qquad (2)$$

Our approach stems from multiple instance learning (MIL), meaning that training set knowledge about specific labels is incomplete (e.g., Speech never directly observed). Here, we model our observed speech data $\mathcal{X}$ as a "bag", containing all co-occurrences of Speech in conjunction with another, possibly noisy background/foreground event label $l \in \{1, \dots, L\}$ from a set of all possible event labels $L < E$ (Equation (2)). So to speak, our approach aims to refine a model's belief about the speech signal $\mathbf{s}$, within complex environmental scenarios. The advantage of this modeling method is that it can be applied for both frame- and clip-level training. Our GPVAD, therefore, relaxes these constraints by allowing training on clip/utterance level, where each training clip contains at least one event of interest. We propose two different models: GPV-F, which outputs $E = 527$ labels ($L = 405$) and the naive GPV-B, $E = 2, L = 1$. GPV-F can be seen as a full-fledged WSSED approach using maximal label supervision and is, therefore, more costly than GPV-B, which only requires knowledge about a clip containing Speech. However, GPV-F should be capable of modeling each individual noise-event instead of clustering all noise into a single class (GPV-B), thus possibly enhancing performance in heavy noise scenarios. The two models are compared against a model trained on frame-level, further referred to as VAD-C.

All models share a common backbone convolutional recurrent neural network (CRNN) [9] approach used in WSSED, which is shown to be robust towards short, sporadic events such as Speech. The following modification to [9] have been done: 1. Add an upsampling operation, such that the models' time-resolution remains constant. 2. Use L^p pooling as our default with $p = 4$, as it has been seen to be beneficial for duration invariant estimates. Different from VAD-C training, where frame-level labels are available, our GPVAD framework is split into two distinct stages. During training, only clip/utterance-level labels are accessible. Therefore a temporal pooling function is required (Equation (4)). During inference, post-processing needs to be applied (Section 4.3) to convert probability sequences into binary labels (absence/presence of an event) as well as any predicted non-speech label is discarded. The framework is depicted in Figure 1.

4. Experiments

In our work, deep neural networks were implemented in PyTorch [17], front-end feature extraction utilized librosa [18]. Code is available online[1].

4.1. Datasets

Datatype	Name	Condition	Label	Duration
Training	Audioset	Real	Clip	15 h
	Aurora 4+	Syn	Frame	30 h
Evaluation	Aurora 4 (A)	Clean	Frame	40 min
	Aurora 4 (B)	Syn	Frame	8.7 h
	DCASE18	Real	Frame	100 min

Table 1: *Training datasets for GPVAD (Audioset) and VAD-C (Aurora 4+), as well as the three proposed evaluation protocols for clean, synthetic noise and real-world scenarios. Duration represents the approximate length of speech.*

Utilized datasets in this work can be split into a train data portion, which differs between the GPVAD and VAD ap-

---

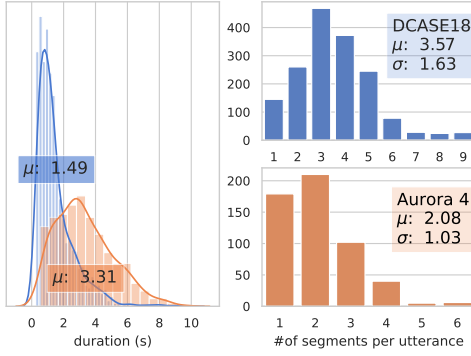[1] Available at github.com/richermans/gpv

Figure 2: *Evaluation data distribution with regards to duration (left) and number of segments per utterance (right), between the Aurora 4 (orange) and DCASE18 (blue) sets. Best viewed in color.*

proaches, and evaluation data, which is shared by both approaches. Our main GPVAD training dataset is the "balanced" set provided by the AudioSet corpus [19], containing 21100/22160 (due to unavailability) 10-second Youtube audio clips, categorized into 527 noisy event labels. From the available 21100 clips (58h), 5452 clips ($\approx$ 15h) are labeled as containing speech, but always alongside $L = 405$ other events (e.g., Bark). Regarding GPV-B, we replace all 526 events in the balanced dataset, not being speech as "noise", thus $\mathcal{X}_{\text{GPV-B}} = \{(\mathbf{s}, \mathbf{u}_{noise},), \mathbf{u}_{noise}\}$. It is important to note that for GPV-B/V training, speech is never individually observed.

Our VAD-C model is trained on the Aurora 4 training set extended by 15 hours of Switchboard [20], obtaining our Aurora 4+ training subset, containing clean as well as synthetic noise data. The additive synthetic noise (Syn) is obtained from six different noise types (car, babble, restaurant, street, airport, and train) that were added at randomly selected SNRs between 10 and 20 dB. All utilized datasets are described in Table 1. Three different evaluation scenarios are proposed. First, we validate on the 40 minutes long, clean Aurora 4 test set [7]. Second, we synthesize a noisy test set based on the clean Aurora 4 test set by randomly adding noise from 100 noise types using a SNR ranging from 5db to 15db in steps of 1db. Lastly, we merge the development and evaluation tracks of the DCASE18 challenge [10], itself a subset of Audioset, to create our real-world evaluation data. The DCASE18 data provides ten domestic environment event labels, of which we neglect all labels other than Speech, but report the number of instances where non-speech labels were present. Our DCASE18 evaluation set encompasses 596 utterances labelled as "Speech", 414 utterances (69%) contain another non-speech label, 114 utterances (20%) only contain speech and 68 utterances (11%) contain two or more non-speech labels.

As it can be seen in Figure 2, the DCASE18 evaluation datasets differ from the Aurora 4 dataset in terms of average duration spoken (1.49 s vs. 3.31 s), as well as number of spoken segments within an utterance (3.87 vs. 2.08).

4.2. Evaluation Metrics

Frame-level For frame-level evaluation, we utilize frame macro/micro averaged F1 scores (F1-macro, F1-micro), Area Under the Curve (AUC) [21], and frame error rate (FER).

Segment-level For segment-level evaluation we utilize event-based F1-Score (Event-F1) [22, 23]. Event-F1 calculates whether onset, offset, and the predicted label overlaps with the ground truth, therefore being a measure for temporal consistency. We set a t-collar value according to WSSED research [10] to 200 ms to allow an onset prediction tolerance and further permit a duration discrepancy between the reference and prediction of 20%.

4.3. Setup

Regarding feature extraction, all experiments used 64-dimensional log-Mel power spectrograms (LMS) in this work. Each LMS sample was extracted by a 2048 point Fourier transform every 20 ms with a window size of 40 ms using a Hann window. During training, zero padding to the longest sample-length within a batch is applied, whereas, during inference, a batch-size of 1 is utilized, meaning no padding.

The training criterion for all experiments between the ground truth $\hat{y}$ and prediction y is cross-entropy Equation (3) for all samples N.

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N} \sum_{n}^{N} \hat{y}_n \log(y_n) + (1 - \hat{y}_n) \log(1 - y_n) \quad (3)$$

Linear softmax [24, 9] (Equation (4)) is utilized as temporal pooling layer that merges frame-level probabilities $y_t(e) \in [0, 1]$ to a single vector representation $y(e) \in [0, 1]^E$.

$$y(e) = \frac{\sum_t^T y_t(e)^2}{\sum_t^T y_t(e)} \quad (4)$$

GPVAD The available training data was split into a label-balanced 90% training and a 10% held-out set for model training using stratification [25]. Due to the inherent label-imbalance within Audioset, sampling is done such that each batch contains evenly distributed clips from each label. Training uses Adam optimization with a starting learning rate of 1e−4, a batch size of 64, and terminates after seven epochs if the criterion did not decrease on the held-out dataset.

VAD-C VAD-C training utilizes a batch size of 20, whereas the loss (Equation (3)) is ignored for padded frames. The learning rate is set to 1e−5, and SGD is used for model optimization. Training target labels are obtained by force alignment from a Kaldi trained ASR HMM model [26].

Post-processing During inference, post-processing is required in order to obtain hard labels from class-wise probability sequences ($y_t(e)$). We hereby use double threshold [9, 14] post-processing, which uses two thresholds $\phi_{\text{low}} = 0.1, \phi_{\text{hi}} = 0.5$.

5. Results

Our results can be seen in Table 2. Firstly, we show that our VAD-C model is capable of performing on an equal footing to other deep neural network approaches [6]. Comparing VAD-C with GPV-B/F, it can be seen that VAD-C is indeed the best performing model given our metrics for clean and synthetic noise datasets. However, evaluation on the real-world dataset reveals a different picture. Here, VAD-C seems to be struggling against the naive GPV-B approach (AUC 87.87 vs. 89.12, FER 21.92 vs. 19.65), indicating that VAD-C is more likely to misclassify speech in the presence of real-world noise. Moreover, in real-world scenarios, GPV-F is outperforming VAD-C for each proposed metric. Our proposed GPV-F approach can also be seen

Testset	Condition	Model	Metric				
			F1-macro(%)	F1-micro(%)	AUC(%)	FER(%)	Event-F1(%)
Aurora 4 (A)	Clean	VAD-C	**96.55**	**97.43**	**99.78**	**2.57**	**78.9**
		GPV-B	86.24	88.41	96.55	11.59	21.00
		GPV-F	<u>95.58</u>	<u>95.96</u>	<u>99.07</u>	<u>4.01</u>	<u>73.70</u>
Aurora 4 (B)	Syn	VAD-C	**85.96**	**90.28**	**97.07**	**9.71**	**47.5**
		GPV-B	73.90	75.75	89.99	24.25	8.0
		GPV-F	<u>81.99</u>	<u>84.26</u>	<u>94.63</u>	<u>15.74</u>	<u>35.4</u>
DCASE18	Real	VAD-C	77.93	<u>78.08</u>	87.87	21.92	<u>34.4</u>
		GPV-B	<u>77.95</u>	75.75	<u>89.12</u>	<u>19.65</u>	24.3
		GPV-F	**83.50**	**84.53**	**91.80**	**15.47**	**44.8**

Table 2: *<mark>Best achieved results</mark> on each respective evaluation condition. Bold marks best result for the respective dataset, while underlined marks second best.*
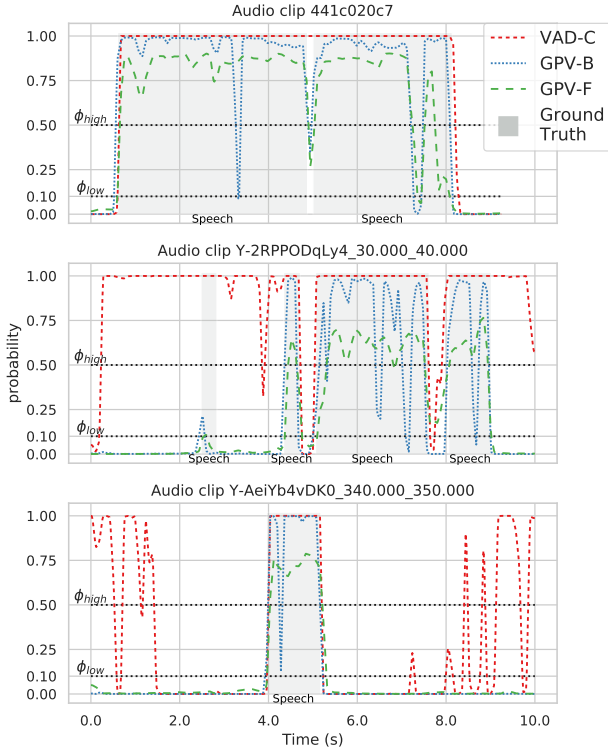

Figure 3: *Per-frame probability output for three sample clips, with visualized speech occurrence (boxed, gray). (Top) Contains a clip from Aurora 4 (B); (Center) contains a musician playing a guitar (DCASE18); (Bottom) contains somebody talking with background noises (DCASE18). Post-processing thresholds ϕ_{high}, ϕ_{low} are indicated. Best viewed in color.*

sampled from the testing set, and per-frame output probabilities are shown for each model seen in Figure 3. In the case of the synthetic Aurora 4 test at the top, we can see that our GPVAD models are capable of modeling short pauses between two speech segments, at which VAD-C fails, yet both GPVAD models could not correctly estimate the second speech segments end. The center sample further demonstrated a typical VAD-C problem in real-world scenarios: it is unable to distinguish between foreground events (here Guitar) and active speech for a majority of the utterance. Especially the bottom sample exemplifies this problem: VAD-C starts to predict speech, where there is none, while both GPVAD models are capable of distinguishing any background noises from speech. Please note that the bottom clip contains laughter at the end, which VAD-C classifies as speech. In our future work, we would like to further extend the scope of GPVAD training by utilizing larger training data (e.g., unbalanced AudioSet).

6. Conclusion

This paper introduces a noise-robust VAD approach by utilizing weakly labeled sound event detection. Two GPVAD systems are investigated: GPV-B, trained on binary speech and non-speech pairs only, as well as GPV-F, which utilizes all 527 AudioSet labels. Our evaluation protocol thoroughly compares our proposed GPVAD approach to traditional VAD utilizing five distinct metrics. Results indicate that GPV-B, even though trained on clip-wise, unconstrained speech, can be used to detect spoken language, without requiring clean, frame-labeled training data. Further, while GPV-B/F both fall short in clean and synthetic noise scenarios against VAD-C, they excel at stable predictions for real-world data. Specifically it can be seen that our proposed approach is robust in its performance across the synthetic and real-world noise datasets. Our best performing model, GPV-F outperforms traditional supervised VAD approaches by a significant margin on real-world data, culminating in an absolute increase of 5.57% F1-macro, 6.45% F1-micro, 3.93% AUC, 6.45% FER and 10.4% Event-F1.

7. Acknowledgements

to be consistently noise-robust since its performance difference between synthetic noise and real-world scenarios is minor.

Even though GPV-B is, on average underperforming against the other two approaches, one should note that it is the least costly system, since labeling data for GPV-B is essentially a binary question whether one heard any speech within a clip, making this approach capable of cheaply scaling to large data. We conclude that the GPVAD models trained with only clip-level labels are capable of competing trained on frame-level labels.

Quantitative Results In order to visualize model-specific behavior, three clips (one Aurora 4 Noisy, two DCASE18) were

8. References

[1] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, oct 2013, pp. 7378–7382.

[2] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks." in *INTERSPEECH*. Lyon, France, 2013, pp. 728–731.

[3] S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, "Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2519–2523.

[4] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-end Domain-Adversarial Voice Activity Detection," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, May 2020.

[5] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 483–487.

[6] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for VAD," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May. Institute of Electrical and Electronics Engineers Inc., may 2016, pp. 5695–5699.

[7] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] H. Dinkel and K. Yu, "Duration Robust Weakly Supervised Sound Event Detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020, pp. 311–315. [Online]. Available: https://ieeexplore.ieee.org/document/9053459/

[10] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 19–23. [Online]. Available: https://hal.inria.fr/hal-01850270

[11] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *arXiv preprint arXiv:1905.10091*, 2019.

[12] Y. Xu, Q. Kong, Q. Huang, W. Wang, and M. Plumbley, "Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging," 08 2017.

[13] E. Çakır, 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, June 2017.

[14] Q. Kong, Y. Xu, I. Sobieraj, W. Wang, and M. D. Plumbley, "Sound Event Detection and Time-Frequency Segmentation from Weakly Labelled Data," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 4, pp. 777–787, apr 2019. [Online]. Available: http://arxiv.org/abs/1804.04715

[15] T. Pellegrini and L. Cances, "Cosine-similarity penalty to discriminate sound classes in weakly-supervised sound event detection," in *International Joint Conference on Neural Networks (IJCNN 2019)*. Budapest, HU: INNS : International Neural Network Society, 2019, pp. 1–8. [Online]. Available: https://oatao.univ-toulouse.fr/24872/

[16] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition," dec 2019. [Online]. Available: http://arxiv.org/abs/1912.10211

[17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[18] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, V. Morozov, R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, Nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss, "librosa/librosa: 0.7.2," jan 2020. [Online]. Available: https://zenodo.org/record/3606573

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 776–780.

[20] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ser. ICASSP'92. USA: IEEE Computer Society, 1992, p. 517–520.

[21] T. Fawcett, "Introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 06 2006.

[22] A. Mesaros, T. Heittola, T. Virtanen, A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for Polyphonic Sound Event Detection," *Applied Sciences*, vol. 6, no. 6, p. 162, may 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[23] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A Framework for the Robust Evaluation of Sound Event Detection," oct 2019. [Online]. Available: http://arxiv.org/abs/1910.08440

[24] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 31–35.

[25] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," *Machine Learning and Knowledge Discovery in Databases*, pp. 145–158, 2011.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.