*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# A Lightweight Framework for Online Voice Activity Detection in the Wild

*Xuenan Xu*[1], *Heinrich Dinkel*[2], *Mengyue Wu*[1†], *Kai Yu*[1†]

[1]MoE Key Lab of Artificial Intelligence
X-LANCE Lab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai, China
[2]Xiaomi Corporation, Beijing, China
{wsntxxn, mengyuewu, kai.yu}@sjtu.edu.cn,Heinrich.dinkel@gmail.com

## Abstract

Voice activity detection (VAD) is an essential pre-processing component for speech-related tasks such as automatic speech recognition (ASR). Traditional VAD systems require strong frame-level supervision for training, inhibiting their performance in real-world test scenarios. Previously, the general-purpose VAD (GPVAD) framework has been proposed to enhance noise robustness significantly. However, GPVAD models are comparatively large and only work for offline evaluation. This work proposes the use of a knowledge distillation framework, where a (large, offline) teacher model provides frame-level supervision to a (light, online) student model. Our experiments verify that our proposed lightweight student models outperform GPVAD on all test sets, including clean, synthetic and real-world scenarios. Our smallest student model only uses 2.2% of the parameters and 15.9% duration cost of our teacher model for inference when evaluated on a Raspberry Pi.

**Index Terms**: Voice activity detection, Sound event detection, teacher-student learning, convolutional recurrent neural networks, lightweight voice activity detection

## 1. Introduction

Voice activity detection (VAD) aims to distinguish speech from non-speech segments in an audio stream [1]. A robust VAD system should be able to differentiate speech segments in audio from non-speech ones, including silence, non-speech human sounds, environmental sounds and all other possible noises in the wild. VAD serves as a crucial pre-processing step for speech and signal processing tasks such as automatic speech recognition (ASR), speaker verification (SV) and text-to-speech synthesis (TTS). Unsupervised VAD was once popular in research [2, 3, 4] due to no requirements for labeled data. With the development of deep learning, deep neural networks (DNN), especially convolutional neural networks (CNN) [5, 6, 7] and recurrent neural networks (RNN) [8, 9, 10] have seen successful applications in VAD. Recent works in VAD endeavor to improve the robustness towards noise and domain mismatch [11, 12, 13, 14, 15] where the training data are noisy datasets synthesized by corrupting clean speech with foreground or background noise as well as end-to-end integration with ASR [16].

Supervised VAD approaches require frame-level labels (the presence of speech in each frame), which are obtained by an alignment given by a hidden Markov model (HMM) trained on clean speech data. Traditional supervised VAD methods are limited by the availability of transcribed ASR data, as well as incapable of being trained on real-world data with unknown noises.

One possible alternative is to manually label the presence of speech in real-world datasets, discarding the ASR pipeline. However, the expensive cost of labor restricts this approach from being used on large datasets. Weakly-supervised training, which only requires clip-level labels, has been recently explored in VAD [17, 18]. Such VAD models trained with clip-level supervision signals are referred to as general-purpose VAD (GPVAD) framework for its robustness towards sounds in the wild.

While the GPVAD framework performs well in real-world evaluation scenarios, its performance on clean and synthetic noise scenarios is sub-par to traditional fully supervised VAD approaches. We hypothesize that this behavior inherently stems not only from the label quality (i.e., incorrect clip-level "speech" labels) of the supervision signal, but rather from its position (no access to frame-level supervision). In order to ameliorate this discrepancy between GPVAD and traditional VAD, we use teacher-student learning to provide frame-level supervision to a student from a weakly-supervised teacher.

Another problem with GPVAD is the parameter redundancy given the fact that lighter weight and shorter inference duration are crucial for pre-processing tasks like VAD. It is shown in previous work [17] that the framework trained with as many as 527 sound event labels (GPV-F) largely outperforms the naive binary classifier. However, it could be possible that detailed knowledge about each noise class is not all necessary since VAD models only need to discriminate speech from non-speech signals. In other words, GPV-F contains redundant parameters for VAD due to the decrease of the target size ($527 \rightarrow 2$). A commonly-used method to distill knowledge from deep models into small models is teacher-student training [19, 20, 21], which prevents small models from under-fitting on large training datasets.

In this paper, we propose using teacher-student learning to develop several lightweight models for real-world VAD applications, amounting to a small footprint (lower than 1 Megabyte on disk). Our experiments verify that our proposed lightweight student models outperform GPVAD on all test sets, including clean, synthetic and real-world scenarios. It should also be noted that the previously used GPVAD framework is an *offline* model, meaning that an entire clip needs to be fed to the model before a prediction can be computed. We eliminate the dependency of output probabilities on future inputs by alternating the architecture, resulting in an *online* GPVAD framework. Our smallest student model only uses 2.2% of the parameters and 15.6% duration cost of our teacher model for inference when evaluated on a Raspberry Pi.

The paper is organized as follows. Section 2 introduces the proposed teacher-student approach. The experimental settings are given in Section 3. In Section 4 results and analysis are presented. Section 5 concludes the paper.

[†] corresponding authors

http://dx.doi.org/10.21437/Interspeech.2021-1977

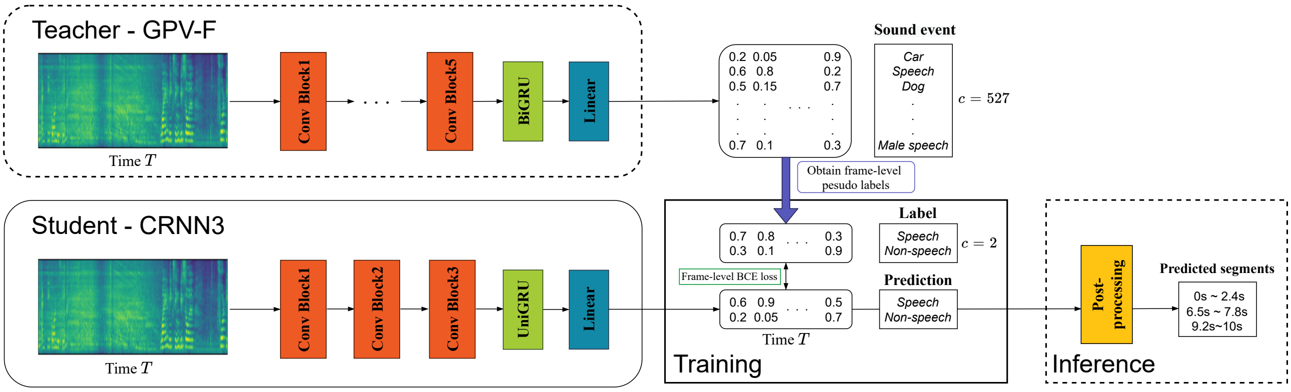

Figure 1: *The proposed Teacher-Student (TS) framework. First, a teacher is trained with clip-level supervision. After training, the teacher provides frame-level supervision to a student model. The knowledge transfer also reduces the amount of learnable labels from 527 to 2. The trained student model is then evaluated.*

## 2. Teacher-student learning with GPVAD

Our teacher-student (TS) framework, depicted in Figure 1, is based on the previously introduced GPVAD paradigm [17]. First, the GPVAD teacher $\mathcal{T}$ model is trained with clip-level supervision on a sound event detection dataset (here AudioSet) to discern between sound events. The teacher has two outputs: a directly trainable clip-level sound event detector and an untrainable frame-level detector. The frame-level sound event detector is indirectly trained via back-propagation from the loss between the clip-level prediction and the ground truth [17, 22, 23].

Then, for a given input audio clip, $\mathcal{T}$ estimates the frame probability $y_t^{\mathcal{T}}(e)$ for each sound event $e$. The estimations are taken as soft-labels to provide frame-level supervision to a student $\mathcal{S}$. Note that $\mathcal{T}$ is trained to predict 527 different events (one of which is speech) while $\mathcal{S}$ is trained as a binary classifier to discriminate speech and non-speech. Therefore, event probabilities predicted by $\mathcal{T}$ are transformed to binary labels for student training.

Since AudioSet contains multiple ambiguous speech-related event labels, student $\mathcal{S}$ is trained on a distilled speech label set $S(\text{Speech})$ containing the parent label "Speech" and all its (seven) children within the AudioSet ontology, e.g., Male speech, Conversation, Child speech. Taking the label set containing all events as $E$, the student training labels $\hat{y}_t^{\mathcal{S}}$ are then defined as:

$$\begin{aligned} S(\text{Speech}) &= \{\text{Speech, Conversation}, \cdots\} \\ \hat{y}_t^{\mathcal{S}}(\text{Speech}) &= \max_{e \in S(\text{Speech})} (y_t^{\mathcal{T}}(e)) \\ \hat{y}_t^{\mathcal{S}}(\text{non-Speech}) &= \max_{e \in E \setminus S(\text{Speech})} (y_t^{\mathcal{T}}(e)) \end{aligned} \tag{1}$$

We use the maximal probability across speech-related events as the representative for the "Speech" class and the maximal probability across all non-speech events as the representative for the "non-Speech" class, since the goal is to teach students to best discriminate speech and non-speech events. Note that, $\hat{y}_t^{\mathcal{S}}(\text{Speech}) + \hat{y}_t^{\mathcal{S}}(\text{non-Speech}) \neq 1$, which enables the student model to predict speech and noise simultaneously. The student is trained by frame-level binary cross entropy (BCE) loss between the prediction $y_t^{\mathcal{S}}$ and ground truth $\hat{y}_t^{\mathcal{S}}$:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} y_t^{\mathcal{S}} \log(\hat{y}_t^{\mathcal{S}}) + (1 - y_t^{\mathcal{S}}) \log(1 - \hat{y}_t^{\mathcal{S}}) \tag{2}$$

During inference, $y_t^{\mathcal{S}}(\text{non-Speech})$ is neglected while only $y_t^{\mathcal{S}}(\text{Speech})$ is taken to predict speech segments.

**Teacher** The GPVAD teacher $\mathcal{T}$ model is a five-layer CRNN model, also known as CDur [23], has achieved competitive performance in SED. CNN recognizes Time-frequency patterns in spectrograms while the bidirectional gated linear unit (BiGRU) is attached to enhance the model's ability to temporally localize sound events. The model architecture can be seen in Table 1.

Table 1: *The detailed configuration of the teacher model. $T$ denotes the frame number of the input feature. Each convolution block contains a batch normalization layer, a 2-dimensional convolution layer and a leaky ReLU activation layer with a negative slope of 0.1. All convolution layers use a $3 \times 3$ filter with zero padding. Each subsampling (Sub) block is denoted as $[t \downarrow f]$, representing a subsampling by factor $t, f$ in time and frequency dimensions respectively. LP-norm subsampling with $p = 4$ is used as default. $c$ represents the number of output labels. All trainable blocks are highlighted in bold.*

| Layer | # Params | Output Size |
|---|---|---|
| **Conv Block1** | 290 | $(32, T, 64)$ |
| Sub1 $[2 \downarrow 4]$ | - | $(32, \frac{T}{2}, 64)$ |
| **Conv Block2** | 36,928 | $(128, \frac{T}{2}, 16)$ |
| **Conv Block3** | 147,712 | $(128, \frac{T}{2}, 16)$ |
| Sub2 $[2 \downarrow 4]$ | - | $(128, \frac{T}{4}, 4)$ |
| **Conv Block4** | 147,712 | $(128, \frac{T}{4}, 4)$ |
| **Conv Block5** | 147,712 | $(128, \frac{T}{4}, 4)$ |
| Sub3 $[1 \downarrow 4]$ | - | $(128, \frac{T}{4}, 1)$ |
| Dropout (0.3) | - | $(128, \frac{T}{4}, 1)$ |
| Reshape | - | $(\frac{T}{4}, 128)$ |
| **BiGRU** | 198,144 | $(\frac{T}{4}, 256)$ |
| **Linear** | 135,439 | $(\frac{T}{4}, c)$ |
| Upsample | - | $(T, c)$ |
| $\sum$ | 813,937 | - |

**Student** While the teacher requires a large number of parameters to sufficiently model 527 sound events, the students are only tasked to learn a binary classification problem, greatly reducing their necessary amount of parameters. In order to decrease the parameter size and enable online evaluation, we focus on two major components. First, since most parameters of the teacher model lie within its convolution layers, we remove the layers containing the most parameters i.e., ultimate and penultimate (4,5) Conv blocks. Second, the bidirectional GRU inhibits online evaluation, since it requires access to future and past input frames. A standard unidirectional GRU replaces it. We therefore propose three model architectures with different channel numbers, denoted as CRNN-C$k, k = 8, 16, 32$, which can be seen in Table 2. Note that an average pooling layer is added after the last convolution to reduce the frequency dimension to one.

Though the number of parameters greatly differs between our teacher and student models, we hypothesize that the resulting student models should be able to perform as well as the teacher in terms of speech prediction.

Table 2: *Parameters of CRNN3 student models. The channel numbers of each convolution block and the GRU hidden unit are listed. A comparison between the parameters sizes against the teacher model is also provided. Note that upsampling is only used during training to match the input time-resolution.*

| | CRNN3-C8 | CRNN3-C16 | CRNN3-C32 |
|---|---|---|---|
| **Conv Block1** | $(8, T, 64)$ | $(16, T, 64)$ | $(32, T, 64)$ |
| Sub1 [2 $\downarrow$ 4] | - | - | - |
| **Conv Block2** | $(32, \frac{T}{2}, 16)$ | $(64, \frac{T}{2}, 16)$ | $(128, \frac{T}{2}, 16)$ |
| Sub2 [2 $\downarrow$ 4] | - | - | - |
| **Conv Block3** | $(32, \frac{T}{4}, 4)$ | $(64, \frac{T}{4}, 4)$ | $(128, \frac{T}{4}, 4)$ |
| Dropout (0.3) | $(32, \frac{T}{4}, 4)$ | $(64, \frac{T}{4}, 4)$ | $(128, \frac{T}{4}, 4)$ |
| Pool Freq | $(32, \frac{T}{4})$ | $(64, \frac{T}{4})$ | $(128, \frac{T}{4})$ |
| Reshape | $(\frac{T}{4}, 32)$ | $(\frac{T}{4}, 64)$ | $(\frac{T}{4}, 128)$ |
| **GRU** | $(\frac{T}{4}, 32)$ | $(\frac{T}{4}, 64)$ | $(\frac{T}{4}, 128)$ |
| **Linear** | $(\frac{T}{4}, 2)$ | $(\frac{T}{4}, 2)$ | $(\frac{T}{4}, 2)$ |
| (Upsample) | $(T, 2)$ | $(T, 2)$ | $(T, 2)$ |
| # Params | 18,076 | 71,476 | 284,260 |
| % Params | 2.2% | 8.7% | 35% |
| Size (kb) | 76 | 284 | 1116 |

## 3. Experiments

**Datasets** In this work, the training dataset is exclusively the balanced subset of AudioSet, identical to the GPVAD training set in our previous work [17]. The dataset contains about 21,000 Youtube audio clips with a maximum duration of 10 seconds. Each audio clip is annotated by one or more event labels from overall 527 sound event categories. Compared with traditional supervised VAD training datasets, this dataset contains unpredictable and unknown, real-world noise. The evaluation datasets are consistent with the previous work [17], including the clean Aurora 4 [24], a synthetically noised Aurora 4 and the real-world DCASE18 dataset [25].

**Feature** In this work, 64-dimensional log mel power spectrograms (LMS) are utilized as audio features using librosa [26]. For each sample, LMS is extracted by a 2048 point short time Fourier transform with a Hann window of 40 ms and 20 ms shift. Since audio clips in AudioSet are variable in duration, all features are padded to the longest sample length in a mini-batch while training. During evaluation and inference, each audio clip is fed to the model independently without padding.

**Training** The dataset is split into a 90% training subset and a 10% validation subset. All student models are trained for at most 300 epochs with an early stop strategy of ten epochs. Training is done using Adam optimization algorithm with a starting learning rate of 1e-3. The neural networks are implemented in the PyTorch [27] framework.

**Post-processing** Post-processing is applied to obtain hard predictions from the output probabilities. In offline VAD, post-processing like double threshold can help smoothen predicted segments and enhance performance [22]. However, in an online setting double threshold cannot be used, thus naive thresholding with a threshold $\phi = 0.3$ is adopted as the default in this work.

**Evaluation Metrics** Following previous work, our model is evaluated from both frame-level and segment-level. For frame-level evaluation, macro and micro F1 scores (F1-macro and F1-micro), area under the curve (AUC) and frame error rate (FER) are adopted. For segment-level evaluation, we choose event-based F1-score [28] (Event-F1), which is commonly utilized in sound event detection evaluation, attaching importance to the prediction accuracy of speech activity onsets and offsets. Disjoint predicted speech segments are penalized by Event-F1. A t-collar of 200 ms is set to allow an onset prediction tolerance. Besides, a 20% duration discrepancy between the reference and the prediction is permitted.

## 4. Results

### 4.1. Teacher student training

We first compare our TS framework performance with the previous weakly-supervised pipeline, the teacher (GPV-F) and strongly-supervised VAD-C [17], listed in Table 3. In this experiment we only focus on the largest student model, CRNN3-C32. Since we change the post-processing method, GPV-F and VAD-C are re-evaluated with naive thresholding. It is shown that the student model significantly outperforms the teacher model on all metrics (e.g., event-F1 56.47 → 72.61 on the clean Aurora 4), indicating the importance of frame-level supervision. The performance gap between weakly- and strongly-supervised models is remarkably reduced in terms of AUC. With proper online post-processing methods, the performance of the GPVAD framework can be further improved. It should be noted that pseudo labels for training inevitably contain more errors compared with traditional HMM-aligned labels. However, the frame-level supervision leads to impressive performance enhancement. We assume that the teacher model learns the pattern of speech occurrence from clip-level labels while the student model's ability to detect speech on- and offsets is inherently improved by frame-level supervision.

To give an intuitive visualization of speech localization performance, we randomly choose two samples from the clean test set Aurora 4. The ground-truth and frame-level probabilities are shown in Figure 2. It is shown that in terms of the speech boundary, CRNN3-C32 performs much better than GPV-F. In the top example, GPV-F predicts two short pauses within the second speech segment ground truth while in the bottom one, the silence segment at around the 8th second predicted by GPV-F is

Table 3: *VAD results of strongly-supervised VAD-C, weakly-supervised GPV-F and the largest student CRNN3-C32. Bold marks the best results for each respective datasets.*

| Testset | Model | Metric | | | | |
|---|---|---|---|---|---|---|
| | | F1-macro (%) | F1-micro (%) | AUC (%) | FER (%) | Event-F1 (%) |
| Clean | VAD-C | **97.20** | **97.88** | **99.79** | **2.12** | **82.35** |
| | GPV-F (Teacher) | 93.84 | 95.10 | 98.83 | 4.90 | 56.47 |
| | CRNN3-C32 | 95.59 | 96.56 | 99.29 | 3.44 | 72.61 |
| Synthetic | VAD-C | **87.59** | **91.02** | **96.90** | **8.98** | **43.99** |
| | GPV-F (Teacher) | 79.52 | 81.52 | 95.11 | 18.48 | 18.98 |
| | CRNN3-C32 | 83.77 | 85.92 | 96.58 | 14.08 | 23.16 |
| Real | VAD-C | 77.03 | 77.20 | 87.15 | 22.80 | 29.7 |
| | GPV-F (Teacher) | 83.66 | 84.77 | 91.63 | 15.23 | 37.66 |
| | CRNN3-C32 | **85.87** | **86.68** | **93.64** | **13.32** | **44.91** |

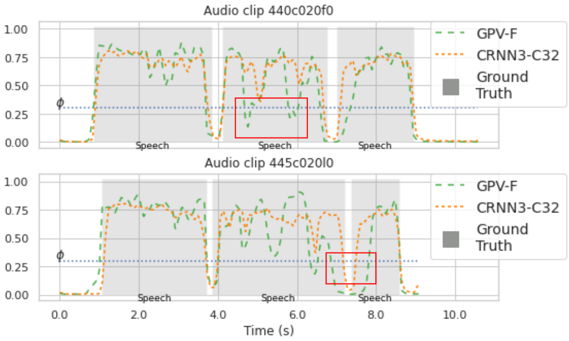

Figure 2: *Frame wise speech probabilities for two clips randomly sampled from Aurora 4. The threshold for post-processing is also depicted. The red boxes indicate the cases where GPV-F fails to give accurate segment boundaries.*

much longer than the ground truth. In comparison, the frame-level trained student model CRNN3-C32 is capable of predicting accurate speech and silence boundaries.

### 4.2. Comparison between student models

Table 4 lists the performance of three student models with different channels. For simplicity, we only list AUC and event-F1 on the real-world test set DCASE18. In teacher-student learning based knowledge distillation, the student performance often degrades with fewer parameters. However, such a phenomenon is not observed in our proposed students: The performance gap between different students is minimal. This validates our assumption that large models contain redundant knowledge on different noise categories. Our lightweight student models distill the knowledge most correlated with speech modeling from the teacher.

Table 4: *Results of different students on the real-world test set DCASE18.*

| Model | AUC (%) | Event-F1 (%) |
|---|---|---|
| CRNN3-C32 | 93.64 | 44.91 |
| CRNN3-C16 | 93.14 | 44.55 |
| CRNN3-C8 | 93.53 | 45.64 |

### 4.3. Model size and inference speed

To compare the computation cost of different models, we test their average inference speed. The test inputs are all 10 second audio clips. The results are presented in Table 5. Our smallest CRNN-C8 only contains 2.7% parameters of GPV-F, with a size of 76 kilobytes on disk. It is lightweight enough to be deployed on embedded systems conveniently, which is crucial for pre-processing techniques like VAD. As Table 5 shows, the inference time on Raspberry Pi is significantly reduced with the decrease of the model size. CRNN3-C8 requires only a 15.6% duration cost of GPV-F for inference, making it capable for low-latency applications.

Table 5: *Comparison of inference speed of different models as well as their floating point operations per second (FLOPS). The inference time is tested on a Raspberry Pi 3 Model B.*

| Model | Raspberry Pi (s) | GFLOPS |
|---|---|---|
| GPV-F | 2.258 | 1.847 |
| CRNN3-C32 | 1.518 | 1.527 |
| CRNN3-C16 | 0.698 | 0.383 |
| CRNN3-C8 | 0.358 | 0.102 |

## 5. Conclusion

In this paper, we propose a teacher-student learning approach to achieve two goals: 1) fill the performance gap on clean and synthetic noise datasets between traditional VAD models and GPVAD by incorporating frame-level supervision; 2) develop a lightweight and online GPVAD framework by knowledge distillation. Three lightweight GPVAD architectures are proposed. Results indicate that teacher-student learning on the same dataset significantly improves VAD performance. The student dramatically surpasses the teacher on all test sets. The largest student CRNN3-C32 achieves an absolute 16.14%, 4.18% and 7.25% event-F1 increase against the teacher on three test sets, respectively. Meanwhile, the model size and computation cost of student models are significantly reduced by knowledge distillation. There is almost no performance degradation brought by model size decrease while the inference on a Raspberry Pi becomes about five times faster. The smallest model occupies only 76 kilobytes on disk, making it suitable for online VAD application in the wild.

# 6. References

[1] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2010.

[2] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[3] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[4] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012.

[5] D. A. Silva, J. A. Stuchi, R. P. V. Violato, and L. G. D. Cuozzo, "Exploring convolutional neural networks for voice activity detection," in *Cognitive Technologies*. Springer, 2017, pp. 37–47.

[6] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5549–5553.

[7] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[8] 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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 483–487.

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

[10] S. Tong, H. Gu, and K. Yu, "A comparative study of robustness of deep learning approaches for vad," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5695–5699.

[11] Y. Lee, J. Min, D. K. Han, and H. Ko, "Spectro-temporal attention-based voice activity detection," *IEEE Signal Processing Letters*, vol. 27, pp. 131–135, 2019.

[12] J. Lee, Y. Jung, and H. Kim, "Dual Attention in Time and Frequency Domain for Voice Activity Detection," in *Proceedings of Conference of the International Speech Communication Association*, 2020, pp. 3670–3674.

[13] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-End Domain-Adversarial Voice Activity Detection," in *Proceedings of Conference of the International Speech Communication Association*, 2020, pp. 3685–3689.

[14] T. Xu, H. Zhang, and X. Zhang, "Polishing the Classical Likelihood Ratio Test by Supervised Learning for Voice Activity Detection," in *Proceedings of Conference of the International Speech Communication Association*, 2020, pp. 3675–3679.

[15] Z. Zheng, J. Wang, N. Cheng, J. Luo, and J. Xiao, "MLNET: An Adaptive Multiple Receptive-Field Attention Neural Network for Voice Activity Detection," in *Proceedings of Conference of the International Speech Communication Association*, 2020, pp. 3695–3699.

[16] T. Yoshimura, T. Hayashi, K. Takeda, and S. Watanabe, "End-to-end automatic speech recognition integrated with ctc-based voice activity detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6999–7003.

[17] Y. Chen, H. Dinkel, M. Wu, and K. Yu, "Voice activity detection in the wild via weakly supervised sound event detection," *Proceedings of Conference of the International Speech Communication Association*, pp. 3665–3669, 2020.

[18] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 1542–1555, 2021.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[20] B. B. Sau and V. N. Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," *arXiv preprint arXiv:1610.09650*, 2016.

[21] M. Huang, Y. You, Z. Chen, Y. Qian, and K. Yu, "Knowledge distillation for sequence model," *Proceedings of Conference of the International Speech Communication Association*, pp. 3703–3707, 2018.

[22] H. Dinkel and K. Yu, "Duration robust weakly supervised sound event detection," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 311–315.

[23] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 29, pp. 887–900, 2021.

[24] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW)*, 2000.

[25] 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 Workshop (DCASE)*, November 2018, pp. 19–23.

[26] B. McFee, V. Lostanlen, M. McVicar *et al.*, "librosa/librosa: 0.7.2," jan 2020.

[27] A. Paszke, S. Gross, F. Massa *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Proceedings of Conference on Neural Information Processing Systems (NIPS)*, 2019, pp. 8026–8037.

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