GPVAD: Towards noise robust voice activity detection via weakly supervised sound event detection

Heinrich Dinkel*, Yefei Chen*, Mengyue Wu and Kai Yu

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
heinrich.dinkel@gmail.com, {chenyefei, mengyuwu, kai.yu}@sjtu.edu.cn

Abstract

Traditional voice activity detection (VAD) methods work well in clean and controlled scenarios, with performance severely degrading in real-world applications. One possible bottleneck for such supervised VAD training is its requirement for clean training data and frame-level labels. In contrast, we propose the GPVAD framework, which can be easily trained from noisy data in a weakly supervised fashion, requiring only clip-level labels. We proposed two GPVAD models, one full (GPV-F), which outputs all possible sound events and one binary (GPV-B), only splitting speech and noise. We evaluate the two GPVAD models and a CRNN based standard VAD model (VAD-C) on three different evaluation protocols (clean, synthetic noise, real). Results show that the GPV-F demonstrates competitive performance in clean and noisy scenarios compared to traditional VAD-C. Interestingly, 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 request for data, the naive binary clip-level GPV-B model can still achieve a 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. Although traditional VAD methods are applied broadly for its high accuracy in clean and controlled environments, its inability to generalize with unseen noise remains unsolved [1]. This inability prevents itself from real-world applications, where existing countless noises with different features and often accompanied by speech. Such scalability limitation might stem from the fact that traditional VAD systems are usually trained on clean/controlled training data, which entitles the models to acquire frame-level labels from a hidden Markov model (HMM). This prerequisite thus leads to constrained use of training data hence refrains VAD systems from training on noisy data.

How VAD systems can generalize with unseen noise inevitably relates to how they can be trained on noisy data. A task to enhance noise robustness while enabling noisy data training, by its nature, is related to weakly-supervised sound event detection (WSSED), which detects and labels all different sound segments, including speech via clip-level supervision. Hence, we will first review VAD and WSSED research, then present our motivation to integrate WSSED methods in scaling VAD to noisy scenarios and relaxing its dependence on frame labeling.

Voice activity detection VAD can be performed either unsupervised [1] with approaches such as energy-thresholding [2], zero-crossing rate [3], or supervised with methods as feature-based and model-based. For feature-based methods, different acoustic features are first extracted, such as time-domain energy [4], zero-crossing rate [3] and pitch [5], then simple detection scheme such as threshold comparison is applied. Regarding model-based methods, separate statistical models are trained to represent speech and non-speech segments by different probability distributions, where the likelihood ratio between the two models is used as a decision threshold. Models such as Gaussian Mixture Model (GMM) [6] and Hidden Markov Model (HMM) [7] have been investigated. Deep learning approaches also have been successfully applied to VAD [8][9][10]. For VAD in complex environments, DNN has better modeling capabilities than traditional methods [9], recurrent neural network (RNN) and long short-term memory (LSTM) can better model long-term dependencies between inputs [11][12][13] and convolutional neural network (CNN) can generate better features for VAD training [10]. However, despite the application of deep learning methods, VAD training still acquires frame labels when possible. Thus training data utilized is usually under controlled environment with or without additional synthetic noise.

Weakly supervised sound event detection Sound event detection (SED) aims to classify (audio tagging) and possibly localize multiple co-occurring sound events (e.g., dog barking) from a given audio clip. Here, 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. Recent advances in weakly supervised sound event detection, in particular, the detection and classification of acoustic scenes and events (DCASE) challenges [14], led to large improvements for predicting accurate sound event boundaries as well as event-labels [15][16][17][18][19][20]. In particular, recent work [21] has shown promising performance regarding short, sporadic events such as speech.

Motivation Since WSSED systems are robust to noise and only require clip-level labels, this work is to investigate two questions: 1) Are current, multi-class WSSED models comparable in performance to VAD? 2) Is utterance-level training a viable alternative compared to frame-level? We thus introduce...
The proposed framework. Two architectures are utilized: GPVAD, trained on weakly-supervised clip-level labels, and VAD trained on fully-supervised frame-level labels. Each Conv2d block represents a batch-normalization, followed by a 2-dimensional convolution with kernel size $3 \times 3$ and a leaky ReLU activation with a negative slope of 0.1. The CNN output is fed into a bidirectional gradient recurrent unit (GRU) with 128 hidden units. The architecture sub-samples the temporal dimension $T$ by a factor of 4 and later upsamples 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 event Speech, is kept for final evaluation.

The advantage of this modeling method is that it can be applied to a wide range of noisy scenarios, including those with short, sporadic events such as speech. The following modification to [21] 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 (5)). During inference, post-processing needs to be applied (Section 3.3) to convert probability sequences into binary labels (absence/presence of an event) as well as any predicted non-speech label is discarded. The framework can be seen in Figure 1.

Figure 1: The proposed framework. Two architectures are utilized: GPVAD, trained on weakly-supervised clip-level labels, and VAD trained on fully-supervised frame-level labels. Each Conv2d block represents a batch-normalization, followed by a 2-dimensional convolution with kernel size $3 \times 3$ and a leaky ReLU activation with a negative slope of 0.1. The CNN output is fed into a bidirectional gradient recurrent unit (GRU) with 128 hidden units. The architecture sub-samples the temporal dimension $T$ by a factor of 4 and later upsamples 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 event Speech, is kept for final evaluation.

2. Approach

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

$$x = s + u$$

However, directly modelling $u$ is rather difficult, since each type of noise has its individual traits. Therefore, we aim at learning the properties of $s$ by observing it with potentially $L$ different non-speech events ($u_1, \ldots, u_L$). Those events are not restricted to being background/foreground noises and can have distinct real-world sounds (e.g., Cat, Music).

$$\mathcal{X} = \{x_1, \ldots, x_i, \ldots, x_L\} \quad (2)$$
$$x_i = (s, u_i) \quad (3)$$

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 $l \in \{1, \ldots, L\}$ 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 $s$, 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, $E = 2, L = 1, \mathcal{X} = (s, u_{\text{Noise}})$. 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. 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) [21] approach used in WSSED, which is shown to be robust towards short, sporadic events such as speech. The following modification to [21] 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 (5)). During inference, post-processing needs to be applied (Section 3.3) to convert probability sequences into binary labels (absence/presence of an event) as well as any predicted non-speech label is discarded. The framework can be seen in Figure 1.

3. Experiments

In our work, deep neural networks were implemented in PyTorch [22], front-end feature extraction utilized librosa [23] and data pre-processing used gnu-parallel [24]. Code will be available online [25].

3.1. Datasets

Utilized datasets in this work can be split into a train data portion, which differs between the GPVAD and VAD approaches, and evaluation data, which is shared by both approaches. Our main GPVAD training dataset is the “balanced” set provided.

---

1. Available at github.com/richermans/gpv
by the AudioSet corpus [25], containing 21100/22160 (due to unavailability) 10-second Youtube audio clips, categorized into 527 webly/noisy event labels. 5452 clips are labeled as containing speech (≈ 15h), but always in tandem with $L = 405$ other events (e.g., Bark) and never occur independently. Regarding GPV-B, we replace any label in the balanced dataset not being speech as “non-speech”. Thus GPV-B is trained on two labels: (Speech, Non-Speech), Non-Speech.

Our VAD-C model is trained on the Aurora 4 training set extended by 15 hours of Switchboard, obtaining our Aurora 4+ training subset, containing clean as well as synthetic noise data. The noise part is obtained by adding 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 minute long, clean Aurora 4 test set [25]. Second, we synthesize a noisy test set based on the clean Aurora 4 test set by randomly adding noise from 100 noise types using an SNR ranging from 5dB to 15dB in steps of 1dB. Lastly, we merge the development and evaluation tracks of the DCASE18 challenge [13], 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 (1.49s vs. 3.31s), as well as number of spoken segments within an utterance (3.87 vs. 2.08).

| Datatype | Name | Condition | Label   | Duration |
|----------|------|-----------|---------|----------|
| Training | Audioset | Real | Clip | 30h |
|          | Aurora 4+ | Noisy | Frame | 30h |
| Evaluation | Aurora 4(A) | Clean | Frame | 8.7h |
|          | DCASE18 | Real | Frame | 100m |

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.

3.2. Evaluation Metrics

Frame-level For frame-level evaluation, we utilize frame macro/micro averaged F1 scores, Area Under the Curve (AUC) [27], and frame error rate (FER).

Segment-level For segment-level evaluation we utilize event-based F1-Score (Event-F1) [28, 29]. Event-F1 calculates whether onset, offset, and the predicted label (Speech) overlaps with the ground truth, therefore being a measure for temporal consistency. We set a t-collar value according to WSSED research [14] to 200ms to allow an onset prediction tolerance and further permit a duration discrepancy between the reference and prediction of 20%.

3.3. Setup

Regarding feature extraction, all experiments used 64-dimensional log-mel power spectrums (LMS) in this work. Each LMS sample was extracted by a 2048 point Fourier transform every 20ms with a window size of 40ms 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.

$$L(\hat{y}, y) = -\frac{1}{N} \sum_{n=0}^{N} \hat{y}_n \log(y_n) + (1 - \hat{y}_n) \log(1 - y_n) \quad (4)$$

The training criterion for all experiments between the ground truth $\hat{y}$ and prediction $y$ is cross-entropy Equation (4) for all samples $N$. Linear softmax [30, 21] (Equation 5) is utilized as temporal pooling layer that merges frame-level probabilities ($y_i(e) \in [0, 1]$) to a single vector representation $y(e) \in \mathbb{R}^P$. 

$$y(e) = \frac{\sum_{t=1}^{T} y_i(e)^2}{\sum_{t=1}^{T} y_i(e)} \quad (5)$$

GPVAD The available training data was split into a label-balanced 90% training and a 10% held-out set for model training using stratification [31]. 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 4) 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.

Post-processing During inference, post-processing is required in order to obtain hard labels from class-wise probability sequences ($y_i(e)$). We hereby use double threshold [21, 18] post-processing, which uses two thresholds $\phi_{low} = 0.1$, $\phi_{hi} = 0.5$.

4. Results

Our results can be seen in Table 2. Firstly, we provide evidence that our VAD-C model is capable of performing on an equal
| Dataset | Model  | F1-macro | F1-micro | AUC   | FER  | F1-Event |
|---------|--------|----------|----------|-------|------|----------|
| 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  | 95.58    | 95.96    | 99.07 | 4.01 | 73.70    |
| Noisy   | VAD-C  | 85.97    | 90.29    | 97.07 | 9.71 | 47.5     |
|         | GPV-B  | 73.90    | 75.75    | 89.99 | 24.25| 8.0      |
|         | GPV-F  | 81.99    | 84.26    | 94.63 | 15.74| 35.4     |
| Real    | VAD-C  | 77.93    | 78.08    | 87.87 | 21.92| 34.4     |
|         | GPV-B  | 77.95    | 75.75    | 89.12 | 19.65| 24.3     |
|         | GPV-F  | 83.50    | 84.53    | 91.80 | 15.47| 44.8     |

Table 2: Achieved best results on each respective evaluation condition. Bold marks best result for the respective dataset, while underlined marks second best.

Quantitative Results

In order to visualize model-specific behavior, three clips (one Aurora 4 Noisy, two DCASE18) were 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 GP-VAD models are capable of modeling short pauses between two speech segments, at which VAD fails, yet both GPVAD models could not correctly estimate the second speech segments end. The center sample further demonstrated a typical VAD problem in real-world scenarios: it is unable to distinguish between foreground events (Guitar) and active speech for a majority of the utterance. Especially the bottom sample exemplifies this problem: VAD 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 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).

5. 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. 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.

6. Acknowledgements

This work has been supported by the Major Program of National Social Science Foundation of China (No.18ZDA293). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.
