An RNN and CRNN Based Approach to Robust Voice Activity Detection

Guan-Bo Wang, Wei-Qiang Zhang
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: wgb14@outlook.com, wqzhang@tsinghua.edu.cn

***Abstract*—In this paper, we propose a voice activity detection (VAD) system, which combines a convolutional recurrent neural network (CRNN) and a recurrent neural network (RNN). In order to improve the performance of our system in low signal-noise ratio conditions, we also add a speech-enhancement module, a one-dimensional dilation-erosion module, and a model ensemble module, all of which contribute significantly. We evaluate our proposed system on development dataset of Public Safety Communications (PSC) and Video Annotation for Speech Technologies (VAST) from NIST Open Speech Analytic Technologies 2019 (OpenSAT19). Compared to the baseline system, our proposed system achieves better performance, using OpenSAT19 official evaluation metrics.**

***Index Terms*—Voice activity detection, RNN, CRNN, speech-enhancement, one-dimensional dilation-erosion, OpenSAT19**

I. Introduction

Voice activity detection (VAD) is now one of the most essential research area, since it is a front-end and contributes in a large number of domains, such as automatic speech recognition, keyword search, speech enhancement, speaker recognition and so on. Traditional approaches of VAD are based on energy spectrum, frequency spectrum, cepstrum, harmonic wave feature, or long-term information. Recently, with significant success and extensive application of deep learning in computer vision and natural language processing, people started to utilize neural networks in VAD. Such as deep belief networks based VAD [1], deep multimodal end-to-end architecture [2] based on both visual network and audio network, diffusion nets based network towards transient noise [3], and frequency-dependent kernel and DIP-based clustering for unsupervised VAD [4].

The National Institute of Standards and Technology (NIST) OpenSAT 2019 is a continuation of the OpenSAT Evaluation Series that started with the 2017 Pilot and organized by NIST. It is planned to have three speech analytics tasks: Speech Activity Detection (SAD), Key Word Search (KWS), and Automatic Speech Recognition (ASR). There are three data domains planned in OpenSAT19: Public Safety Communications (PSC), Video Annotation for Speech Technologies (VAST), and Low Resourced Language (LRL). Our proposed system aims at SAD task in PSC and VAST domains [5].

The corresponding author is Wei-Qiang Zhang.

This work was supported by the National Natural Science Foundation of China under Grant No.U1836219 and the National Key R&D Program of China.

In our paper, we propose a neural network architecture for voice activity detection in low signal-noise ratio situation. Firstly, we extract the log Mel-scale Filter Bank energies (fbank) feature of each frame of the audio. Then we train a convolutional recurrent neural network (CRNN) and a recurrent neural network (RNN) that output whether there exists speech in every frame. We also add a speech-enhancement module, a one-dimensional dilation-erosion module, and a model ensemble module to our system, which greatly improve the final performance.

We evaluate our proposed system on dataset of PSC and VAST [6], both from NIST Open Speech Analytic Technologies 2019 (OpenSAT19). Our evaluation metrics are false positive rate (P_{FP}), false negative rate (P_{FN}) and detection cost function value ($DCF(\theta)$). A false positive (FP) is detecting speech where there is no speech, also called a "false alarm". A false negative (FN) is missed detection of speech, i.e., not detecting speech where there is speech, also called a "miss". The $DCF(\theta)$ is the detection cost function value for a system at a given system decision-threshold setting. The evaluation metrics mentioned above are computed as follows:

$$P_{\mathrm{FP}} = \frac{\text{total FP time}}{\text{annotated total nonspeech time}} \tag{1}$$

$$P_{\mathrm{FN}} = \frac{\text{total FN time}}{\text{annotated total speech time}} \tag{2}$$

$$DCF(\theta) = 0.75 \times P_{\mathrm{FN}} + 0.25 \times P_{\mathrm{FP}} \tag{3}$$

where θ denotes a given system decision-threshold setting. Our goal is to minimize the $DCF(\theta)$ in (3).

The rest of this paper is organized as follows. In section II, we introduce our methods in detail, mainly including system overview, feature extraction, neural network, speech-enhancement, one-dimensional dilation-erosion algorithm, and model ensemble. The dataset, experiments and results are presented in section III. Finally, we conclude our work in section IV.

II. Methods

A. System Overview

Our proposed system is shown in Fig.1. Firstly, input audios are enhanced by speech-enhancement module. Next, we extract fbank features of raw audios and enhanced audios, and then put the features into both a convolutional recurrent neural

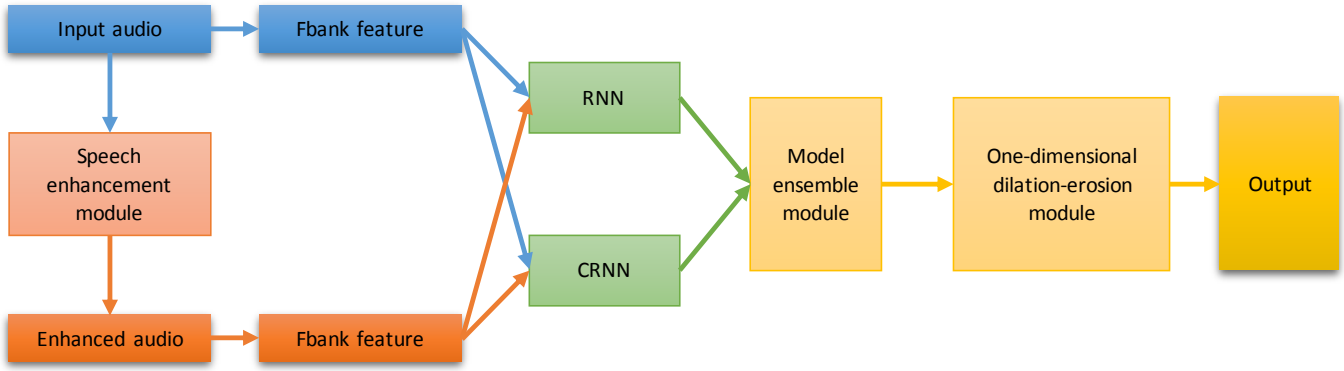

Fig. 1. Overall architecture of proposed system.

network (CRNN) and a recurrent neural network (RNN). The two networks will output speech existence in every frame separately. Finally, we pass the output to the model ensemble module and one-dimensional dilation-erosion module and get the final output.

B. Feature Extraction

In this part of the system, we follow the work in [7]. To imitate non-linear response to the sound spectrum of human ear, we choose the log Mel-scale Filter Bank energies (fbank) feature. First, we convert sampling rate of input audios to 16 kHz, and divide each audio into frames. The frame length is 20 ms and the frame shift is 10 ms. Then we extract fbank feature of each frame, applying 40 mel-scale filters on the magnitude spectrum, which cover the entire range from 0 to 8000 Hz. After that, we take logarithm on the amplitude and then get the fbank feature. Finally, in order to be prepared to be fed into neural networks, the extracted feature is normalized to zero mean and unit standard deviation.

C. Neural Network

In this part of our proposed system, we train two different neural networks, a CRNN and an RNN. Both networks are shown in Fig.2. We cut input feature into 64-frame length segments, so the input shape is 64×40. In CRNN structure, the first two layers are 2-dimensional convolution (Conv-2D) layers, and the next two layers are bi-directional gated recurrent unit (bi-GRU) layers. The final layer is a fully-connected layer, which outputs whether speech exists in each frame. In RNN structure, there are 4 bi-GRU layers after the input layer, and the final layer is a fully-connected layer, which outputs whether speech exists in each frame.

D. Speech-Enhancement

In this module, our work is based on a regression approach to speech enhancement based on deep neural networks (DNN) [8]. To train the speech-enhancement model, we utilize the same datasets, the same neural network, and the same training method mentioned or provided in [8]. We add noise audio

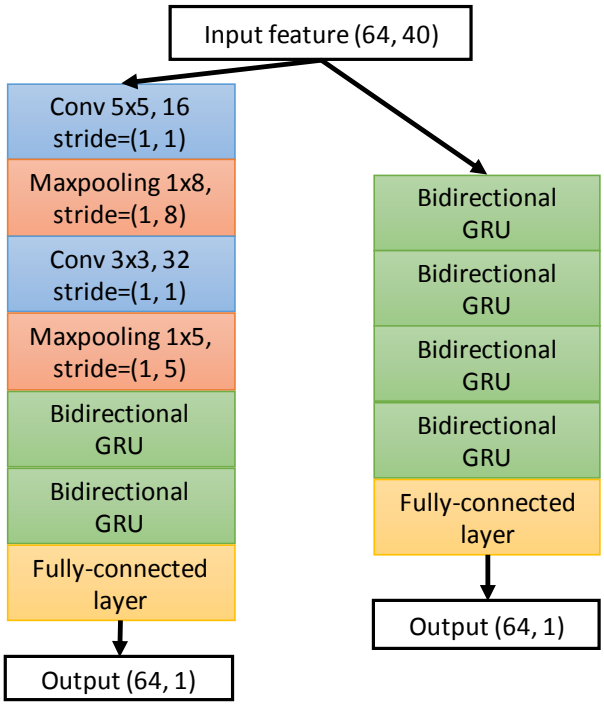

Fig. 2. The architecture of CRNN and RNN.

to clean speech audio and get noisy speech. Then both noisy speech and clean speech are fed into the DNN as input and output of the network separately. In this way, we get a DNN based speech-enhancement module. We also listen to the output of the module, and although the speech enhancement results don't sound perfect, it does improve the VAD performance.

E. One-Dimensional Dilation-Erosion

Due to the defect of speech-enhancement, speech is inevitably weakened when background noise is reduced. Therefore, some "holes" that appear in speech segments in output

need to be patched, and some boundaries need to be expanded. The dilation-erosion algorithm is widely used in computer vision area, which means dilating or eroding white area on binary image. In our system output, each audio is divided into several segments marked speech or nonspeech, and this is just like a one-dimensional binary image. Therefore, our one-dimensional dilation-erosion algorithm is dilating or eroding speech segments of output. The specific algorithm is implemented as Algorithm 1.

Algorithm 1 One-Dimensional Dilation-Erosion Algorithm

Input: $mode$, $bias$
  for all $segment$ **do**
    if $mode = dilation$ **then**
      if $segment\_state = speech$ **then**
        $segment\_start\_time \Leftarrow segment\_start\_time - bias$
        $segment\_end\_time \Leftarrow segment\_end\_time + bias$
      else
        $//\ segment\_state = nonspeech$
        $segment\_start\_time \Leftarrow segment\_start\_time + bias$
        $segment\_end\_time \Leftarrow segment\_end\_time - bias$
      end if
    else
      $//\ mode = erosion$
      if $segment\_state = speech$ **then**
        $segment\_start\_time \Leftarrow segment\_start\_time + bias$
        $segment\_end\_time \Leftarrow segment\_end\_time - bias$
      else
        $//\ segment\_state = nonspeech$
        $segment\_start\_time \Leftarrow segment\_start\_time - bias$
        $segment\_end\_time \Leftarrow segment\_end\_time + bias$
      end if
    end if
  end for
  for all $segment$ **do**
    if $segment\_start\_time > segment\_end\_time$ **then**
      delete $segment$
    end if
  end for

F. Model Ensemble

In order to get better performance, we add model ensemble module to our proposed VAD system, a commonly applied strategy. We choose the BUT phoneme recognizer [9] as our ensemble model, which is also our baseline system. The method for model ensemble is WCombSum algorithm [10] used in keyword search (KWS) area. WCombSum is computed as follows:

$$s = \sum_{i=1}^{N} w_i \times s_i \tag{4}$$

where w_i denotes weight of each system, and s_i denotes confidence score for each system. We normalize the output of each system into a range of 0 to 1, so that we can find the optimal threshold θ within a certain range to minimize the $DCF(\theta)$.

III. Experiments and Results

A. Dataset

We use a total of four datasets, PSC training data set (PSC_train), PSC development data set (PSC_dev), VAST development data set in 2017 (17VAST_dev), and VAST development data set (VAST_dev), three of them provided in OpenSAT19 and one provided in OpenSAT17. Each dataset mentioned above has its own transcripts. There are 131.2 hours of audio in PSC_train set, 5 hours in PSC_dev, 13.3 hours in 17VAST_dev, and 13.5 hours in VAST_dev, including speeches with no background, quiet background, loud background or babble.

B. Experiment

We choose PSC_train set as our training set, and PSC_dev set as well as VAST_dev set as our test set.

During the training phase, we feed acoustic feature and transcripts of PSC_train set into our CRNN and RNN models. Batch normalization [11] and dropout [12] are used in training phase, which are not shown in Fig.2. Our models are trained using Adam optimizer [13], with initial learning rate 0.001. Binary cross-entropy is used as the loss function and the batch size is 256. We randomly select 20% of training set as validation set. The training will be stopped after 100 epochs, and we also use early stopping strategy when the validation loss stops degrading for 10 epochs. Our evaluation metric is the $DCF(\theta)$ mentioned in section I.

When it comes to the test phase, we put the audio in test set into speech enhancement module, and get enhanced audio. Features of both raw audio and enhanced audio are put into our trained models. The outputs of neural networks and BUT phoneme recognizer will pass model ensemble module and one-dimensional dilation-erosion module. In the end, we get the final output of our proposed system. We evaluate the system using the $DCF(\theta)$, with a collar of 500 ms between each speech and nonspeech.

Additionally, we also feed audio feature and transcripts of 17VAST_dev set into our neural networks when evaluating the performance in VAST_dev set. This does work effectively, because the 17VAST_dev set can supplement the noise information of the VAST set, which is not available in the PSC set. Noticing the data imbalance in VAST set, we use weighted cross-entropy loss function computed as follows:

$$Loss = -\frac{1}{N}\sum w_S \hat{y}_t \log(y_t) + w_N(1-\hat{y}_t)\log(1-y_t) \tag{5}$$

where y_t denotes the output of each frame, $\hat{y}_t$ denotes the ground-truth label, and w_S and w_N denote the weights of speech and nonspeech frames, respectively.

TABLE I
PERFORMANCE OF SPEECH-ENHANCEMENT MODULE AND ONE-DIMENSIONAL DILATION-EROSION MODULE. *** INDICATES THAT THE CORRESPONDING RESULT IS NOT AVAILABLE AND – INDICATES THAT THE CORRESPONDING RESULT IS MEANINGLESS. (1)OFFICIAL BASELINE: BASELINE OUTPUT PROVIDED BY OPENSAT17; (2)RAW: BUT; (3)RAW_DE: BUT WITH ONE-DIMENSIONAL DILATION-EROSION; (4)SE: BUT WITH SPEECH-ENHANCEMENT; (5)SE_DE: BUT WITH SPEECH-ENHANCEMENT AND ONE-DIMENSIONAL DILATION-EROSION.

system	PSC_dev			VAST_dev			17VAST_dev			
	P_{FN}	P_{FP}	DCF	P_{FN}	P_{FP}	DCF	P_{FN}	P_{FP}	DCF	Relative Improvement(%)
Official Baseline	***	***	***	***	***	***	0.3606	0.0726	0.2886	–
Raw	0.0735	0.7442	0.2412	0.0533	0.7609	0.2302	0.0199	0.8070	0.2167	24.9
Raw_DE	0.0058	0.8507	0.217	0.0011	0.8570	0.2150	0.0005	0.8646	0.2165	25.0
SE	0.1897	0.4825	0.2629	0.2311	0.4273	0.2802	0.1514	0.5009	0.2388	17.2
SE_DE	0.0216	0.6169	0.1705	0.0258	0.6252	0.1757	0.0190	0.6887	0.1864	35.4

C. Results

Tab.I shows the performance of speech-enhancement module and one-dimensional dilation-erosion module. Note that the official baseline is provided in OpenSAT17, actually not our baseline. As shown in Tab.I, compared to using BUT phoneme recognizer simply, utilizing both our speech-enhancement and one-dimensional dilation-erosion algorithm will reduce the $DCF(\theta)$ by 0.0707, 0.0545, and 0.0303, relatively improving 29.3%, 23.6%, and 14.0% in three data sets, respectively.

TABLE II
PERFORMANCE OF OUR PROPOSED SYSTEM, * INDICATES THE SYSTEM USING SPEECH-ENHANCEMENT MODULE AND ONE-DIMENSIONAL DILATION-EROSION MODULE

system	PSC_dev			VAST_dev		
	P_{FN}	P_{FP}	DCF	P_{FN}	P_{FP}	DCF
BUT	0.0735	0.7442	0.2412	0.0533	0.7609	0.2302
BUT*	0.0216	0.6169	0.1705	0.0258	0.6252	0.1757
CRNN	0.0998	0.3585	0.1644	0.0805	0.3399	0.1453
CRNN*	0.0789	0.2651	0.1255	0.0517	0.1997	0.0887
RNN	0.1199	0.2880	0.1619	0.0863	0.3164	0.1438
RNN*	0.0808	0.1889	0.1078	0.0835	0.1720	0.1056
Ensemble	0.1498	0.1063	0.1389	0.0643	0.3437	0.1342
Proposed	0.0604	0.0472	**0.0571**	0.0481	0.1944	**0.0846**

The final performance in PSC_dev set and VAST_dev set of our proposed system is shown in Tab.II. Finally, our system achieves the lowest $DCF(\theta)$ both in PSC_dev set (0.0571) and VAST_dev set (0.0846), outperforming other methods. Compared to baseline performance, we get a 76.3% and a 63.2% improvement in PSC_dev set and VAST_dev set, respectively. Although not tested in the same datasets, our proposed system achieves comparable performance to the top systems in OpenSAT17 [14].

IV. CONCLUSIONS

In this paper, we have introduced an RNN and CRNN based approach to robust voice activity detection. To be more specific, we combine a CRNN model and an RNN model, and then add a speech-enhancement module, a one-dimensional dilation-erosion module, and a model ensemble module. Experimental results show that all of the modules contribute significantly, with P_{FP}, P_{FN} and $DCF(\theta)$ decreased evidently. Compared to the baseline system, our proposed system outperforms greatly and achieves comparable performance to the top systems in OpenSAT17. We believe the proposed system will become a helpful front-end of ASR and KWS systems, and will contribute in various domains. As for performance in other datasets, more future work needs to be done for futher evaluation.

REFERENCES

[1] 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.

[2] I. Ariav and I. Cohen, "An end-to-end multimodal voice activity detection using wavenet encoder and residual networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 265–274, 2019.

[3] A. Ivry, B. Berdugo, and I. Cohen, "Voice activity detection for transient noisy environment based on diffusion nets," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 254–264, 2019.

[4] H. Dubey, A. Sangwan, and J. H. Hansen, "Leveraging frequency-dependent kernel and dip-based clustering for robust speech activity detection in naturalistic audio streams," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2056–2071, 2018.

[5] "NIST open speech analytic technologies 2019 (OpenSAT19)," https://sat.nist.gov/, accessed June, 2019.

[6] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[7] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, "Learning how to listen: A temporal-frequential attention model for sound event detection," in *Interspeech*, Graz, Austria, Sept 15-19 2019.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[9] P. Ace, P. Schwarz, and V. Ace, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Citeseer, 2009.

[10] J. Mamou, J. Cui, X. Cui, M. J. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran *et al.*, "System combination and score normalization for spoken term detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8272–8276.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[12] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, PhD thesis, University of Cambridge, 2016.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[14] F. R. Byers and S. O. Sadjadi, "2017 pilot open speech analytic technologies evaluation (2017 NIST Pilot OpenSAT) post evaluation summary," Tech. Rep., 2019.