Joint Training ResCNN-based Voice Activity Detection with Speech Enhancement

Tianjiao Xu, Hui Zhang* and Xueliang Zhang
Department of Computer Science, Inner Mongolia University, Hohhot, China
E-mail: xtj@mail.imu.edu.cn alzhu.san@163.com cszxl@imu.edu.cn
Tel/Fax: +86-0471-4993132

***Abstract*—Voice activity detection (VAD) is considered as a solved problem in noise-free condition, but it is still a challenging task in low signal-to-noise ratio (SNR) noisy conditions. Intuitively, reducing noise will improve the VAD. Therefore, in this study, we introduce a speech enhancement module to reduce noise. Specifically, a convolutional recurrent neural network (CRN) based encoder-decoder speech enhancement module is trained to reduce noise. Then the low-dimensional features code from its encoder together with the raw spectrum of noisy speech are feed into a deep residual convolutional neural network (ResCNN) based VAD module. The speech enhancement and VAD modules are connected and trained jointly. To balance the training speed of the two modules, an empirical dynamic gradient balance strategy is proposed. Experimental results show that the proposed joint-training method has obvious advantages in generalization ability.**

I. Introduction

Voice activity detection (VAD) is widely used in many applications, such as speaker recognition, voice wake-up and automatic speech recognition (ASR). The task of VAD is to identify speech or non-speech events in a given audio signal.

At present, VAD is usually treated as a binary-classification problem with pre-marked labels on each frame. The state-of-the-art algorithms are usually based on deep learning architectures, e.g. fully connected deep neural networks (DNNs) [1], [2], convolutional neural networks (CNNs) [3], [4], long short-term memory (LSTM) recurrent neural networks [5]–[7]. Deep learning based methods demonstrated excellent performance in trained noisy conditions. While their performance will drop down in untrained noisy conditions. Improving the performance both in the trained and untrained conditions is improving the discriminative ability and generalization ability, which is an important issue in the supervised learning.

Despite the ongoing development over the years, VAD is still a challenging task in low signal-to-noise ratio (SNR), especially under unmatched noisy conditions. To cope with this challenge, a speech enhancement module is involved. In [8], a multi-targets learning framework is employed, where VAD is jointly trained with a speech enhancement task by hard sharing of underlying parameters to achieve robustness under noisy conditions. A more intuitively method is using a speech enhancement module to reduce noise and get “clean features”, then utilize the VAD module to discriminate speech or non-speech events based on the clean features. Training these two modules jointly has shown excellent improvement [9]. The potential of speech enhancement module is investigated further in [10], where the input features of VAD is not only the clean features but also the intermediate representation from the denoising variational autoencoders (DVAE) based speech enhancement module. However, the performance of these approaches are highly dependent on the performance of the speech enhancement part, which is more difficult than VAD obviously.

In fact, the target of speech enhancement and VAD are somewhat consistent. VAD aims to detect when speech appears in time dimension, and speech enhancement aims to separate the target speech from background in both of time and frequency dimensions. Since the speech enhancement requires higher precision, many speech processing systems employ VAD as a pre-processing step of speech enhancement. It is more difficult to perform speech enhancement excellent in low SNR under untrained noisy conditions. Applying speech enhancement to get “clean features” and improve the performance of VAD is fails when the speech enhancement module is poor.

With these observations, the clean features are abandoned. In consideration of VAD as a pre-processing step of speech enhancement, we speculate that the underlying features of the speech enhancement model may hide useful and valid information for VAD. So an encoder-decoder architecture is employed, which is suitable for underlying features extraction. Specifically, we use a convolutional recurrent neural network (CRN) for speech enhancement, which is encoder-decoder and is effective for speech enhancement tasks [11]. In the proposed method the underlying features from the encoder is concatenated to the feature of noisy speech then feed into a deep residual convolutional neural network (ResCNN) based VAD module. The speech enhancement and VAD modules are connected and trained jointly with an empirical dynamic gradient balance strategy to balance the training speed of the two modules. Experimental results show that the proposed approach outperforms the conventional methods, and helpful to address the issue of performance degradation due to untrained noise.

The rest of this paper is constructed as follows. Section 2 reviews the speech enhancement and VAD then introduces our joint training approach and describes the proposed architecture. Then the experimental details and results are described in section 3. Finally, section 4 concludes the paper.

II. Joint Training VAD

A. Speech Enhancement

Speech enhancement aims to separate target speech from the background interference. Recently, speech enhancement has been formulated as a supervised learning, which employs deep neural network to map from noisy speech to target speech. The commonly used acoustic feature is magnitude spectrum. The input is the magnitude spectrum of noisy speech, and the output is the magnitude spectrum of clean speech [12]–[14]. The mean squared error (MSE) is usually used as the loss function:

$$L_{ss} = \frac{1}{N} \sum_{t=1}^{N} \left(\sum_{f=1}^{M} \left\| Y_{(t,f)} - \hat{Y}_{(t,f)} \right\|^2 \right) \tag{1}$$

where N is the number of frames, M is the number of frequency bins. $Y_{t,f}$ and $\hat{Y}_{t,f}$ represent the feature of clean speech and the estimated speech in the (t, f) time-frequency unit, respectively.

Good speech enhancement can remove the interference noise and then improve VAD. But when the performance of speech enhancement module is poor, especially in low SNR under untrained noisy conditions, it may harm rather than help the VAD. Instead of using the enhancement results which may include much errors, we back to the middle way of the whole speech enhancement processing, which may partly solve the speech enhancement task but does not involve many errors specific to the speech enhancement.

Specifically, we use a CRN [11] to address the speech enhancement. The CRN is an encoder-decoder structure which is shown in the higher part of Fig. 1. A series of convolutional layer is located in the bottom layer of CRN, which can reduce spectral variations effectively. In the intermediate layer, long short-term memory (LSTM) can model the sequential information of the speech signal, which is a conventional choice for both VAD and speech enhancement [15], [16]. After that, a series of deconvolutional layer is employed and made the whole network optimized as an encoder-decoder structure by using noisy-clean speech pairs. More over, the skip layer connection is applied to connect each encoder layer to the corresponding decoder layer, which can improve the flow of gradients when the network is deep and promote the intermediate LSTM layer to be more robust.

B. Voice Activity Detection

VAD aims to detect the speech or non-speech events in an audio signal, which is pretty simple for clean speech. However, for noisy speech, especially in low SNR scenario, VAD is a challenge. To cope with the challenge, in recent years, most researches focused on deep learning methods, which treat VAD as a binary-classification problem and train model on pre-marked corpora. The binary cross-entropy is usually used as the loss function:

$$L_{vad} = -\sum_{t=1}^{N} \left(Y_t \log \hat{Y}_t + (1 - Y_t) \log(1 - \hat{Y}_t) \right) \tag{2}$$

where N is the number of frame, Y_t and $\hat{Y}_t$ represent the VAD label and the estimated label of the t-th frame, respectively.

TABLE I
Architecture of the proposed joint training method. Here T denotes the number of time frames in the magnitude spectrum.

Component	Layer	Hyperparameters	Output Size
Speech Enhancement	Reshape_1	-	$T \times 161 \times 1$
	Conv2d_1	$1 \times 3, (1, 2), 8$	$T \times 80 \times 8$
	Conv2d_2	$1 \times 3, (1, 2), 8$	$T \times 39 \times 8$
	Conv2d_3	$1 \times 3, (1, 2), 8$	$T \times 19 \times 8$
	Conv2d_4	$1 \times 3, (1, 2), 16$	$T \times 9 \times 16$
	Conv2d_5	$1 \times 3, (1, 2), 16$	$T \times 4 \times 16$
	Reshape_2	-	$T \times 64$
	Lstm_1	64	$T \times 64$
	Lstm_2	64	$T \times 64$
	Deconv2d_1	$1 \times 3, (1, 2), 16$	$T \times 9 \times 16$
	Deconv2d_2	$1 \times 3, (1, 2), 8$	$T \times 19 \times 8$
	Deconv2d_3	$1 \times 3, (1, 2), 8$	$T \times 39 \times 8$
	Deconv2d_4	$1 \times 3, (1, 2), 8$	$T \times 80 \times 8$
	Deconv2d_5	$1 \times 3, (1, 2), 1$	$T \times 161 \times 1$
Voice Activity Detection	Reshape_1	-	$T \times 225 \times 1$
	Conv2d_1	$5 \times 5, (1 \times 2), 16$	$T \times 113 \times 16$
	Res_1	$[3 \times 3, 16]$ $[3 \times 3, 16]$	$T \times 113 \times 16$
	Conv2d_2	$5 \times 5, (1 \times 2), 32$	$T \times 57 \times 32$
	Res_2	$[3 \times 3, 32]$ $[3 \times 3, 32]$	$T \times 57 \times 32$
	Conv2d_3	$5 \times 5, (1 \times 2), 32$	$T \times 29 \times 32$
	Res_3	$[3 \times 3, 32]$ $[3 \times 3, 32]$	$T \times 29 \times 32$
	Conv2d_4	$5 \times 5, (1 \times 2), 16$	$T \times 15 \times 16$
	Res_4	$[3 \times 3, 16]$ $[3 \times 3, 16]$	$T \times 15 \times 16$
	Reshape_2	-	$T \times 240$
	FC	1	$T \times 1$

Convolutional neural networks (CNNs) have been proved effective in VAD [17]. The weights sharing technology makes CNN can build a large model with few trainable parameters. Dilating and gating improve the CNNs' performance further, which modeling the temporal sequence better than recurrent neural networks (RNNs) [3], [4]. A skip connection creating a shortcut in a sequential network effectively, which helps prevent information loss along the data-processing flow by adding a past output tensor to a later output tensor. Utilizing residual connection improves the performance of CNN a lot [4], [18]. Therefore we use ResCNN for VAD in this study, which is shown in the lower part of Fig. 1.

C. Joint Training

The joint training approach for VAD with speech enhancement was first introduced in [9], which pre-trained a DNN to map the noisy to clean speech features firstly, then apply a DNN-based VAD to discriminate speech against noise backgrounds and optimized them jointly. In [10], researchers proposed a denoising variational autoencoders (DVAE) for

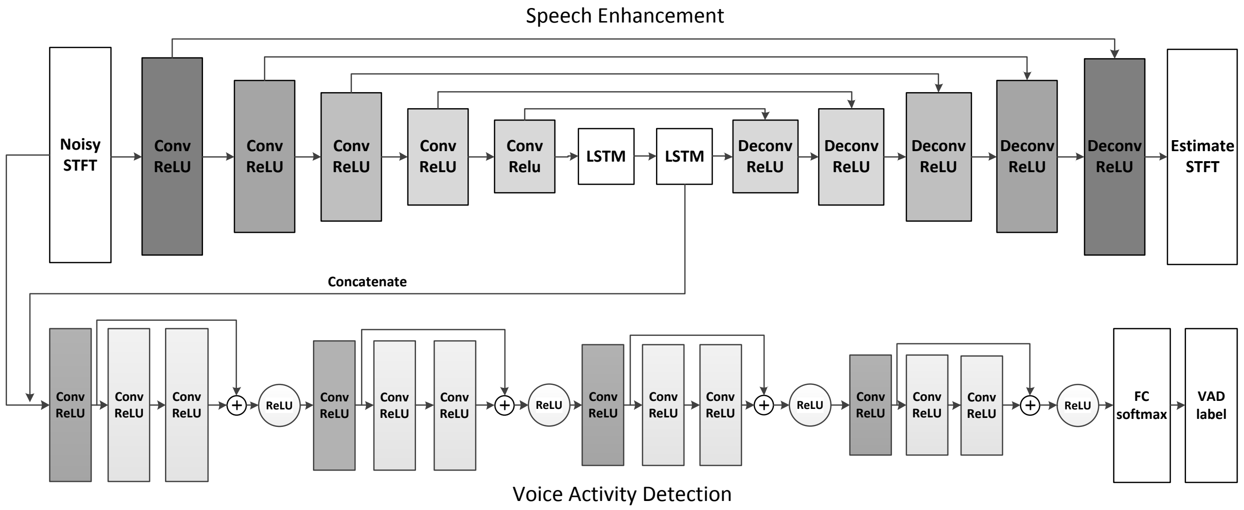

Fig. 1. Network architecture of the proposed joint training.

VAD, which combine the denoised feature and the hidden variable from DVAE and effectively train the whole network without pre-training.

For pre-train methods, the VAD is addressed using estimated speech feature from a fully optimized speech enhancement module. For the non-pre-trained methods, there usually applied a hyper-parameter α to balance the trade-off between the speech enhancement and VAD tasks. So the joint training loss function can be defined as below:

$$L_{jt} = \alpha L_{ss} + (1-\alpha) L_{vad}, \alpha \in [0,1] \tag{3}$$

where L_{ss} represents the loss function of the speech enhancement (Eq. 1). L_{vad} represent the loss function of the VAD (Eq. 2).

However, tuning this hyper-parameter by hand is a difficult and expensive process. With a mountain of experiments, we found that when the gradient descent rate of each task is close to each other, the learning curve tends to be smoother and more stable, and will convergence better than a fixed value. So we proposed a strategy to balance the trade-off between speech enhancement and VAD by adjusting the α.

We adjust α based on two epochs of historical loss value $L_{ss}^{(i-1)}$, $L_{ss}^{(i)}$, $L_{vad}^{(i-1)}$, $L_{vad}^{(i)}$ which is the loss function values of speech enhancement and VAD task at epoch $i-1$, and i. We expect the rate of change or gradient of L_{ss} and L_{vad} is close to each other, which means $C_{ss}^{(i)} = \frac{\left|L_{ss}^{(i)} - L_{ss}^{(i-1)}\right|}{L_{ss}^{(i-1)}}$ should close to $C_{vad}^{(i)} = \frac{\left|L_{vad}^{(i)} - L_{vad}^{(i-1)}\right|}{L_{vad}^{(i-1)}}$. An absolute value is take here to avoid obtaining same sign but in opposite direction. To avoid adjusting α too often, we cache the $C_{ss}^{(t)}$ and $C_{vad}^{(t)}$ two epochs and apply the adjusting strategy when both epochs give much different gradient of loss function. The proposed adjusting strategy is summarized in Algorithm 1. Specifically, we start with an initial parameter α_0. This procedure is iterated until the loss function is fully optimized, which reached the maximal training epoch I, for example.

Algorithm 1: Gradient Balance

```
Input: initial parameter $\alpha_0$
for $i = 1, 2, 3, ..., I$ do
    Compute $L_{ss}^{(i)}$ and $L_{vad}^{(i)}$
    if $i > 1$ then
        Compute $C_{ss}^{(i)} = |L_{ss}^{(i)} - L_{ss}^{(i-1)}|/L_{ss}^{(i-1)}$
        Compute $C_{vad}^{(i)} = |L_{vad}^{(i)} - L_{vad}^{(i-1)}|/L_{vad}^{(i-1)}$
        Compute $M^{(i)} = C_{vad}^{(i)} - C_{ss}^{(i)}$
        if $i > 2$ and $M^{(i)} * M^{(i-1)} > 0$ then
            Set
            $\alpha \leftarrow Max\{Min\{\alpha + (M^{(i)} + M^{(i-1)}), 1\}, 0\}$
        end
    end
end
```

D. Network Architecture

In this study, we construct the speech enhancement model based on CRN which is shown in Fig. 1. The input feature is encoded by 5 layers of 2-D convolutional, which increase the number of channels while reducing the size of the feature map. Then a 64-dimensional sequence of feature vectors are modeled by two LSTM layers with 64 cells. Subsequently, the output sequence of the LSTM layers is decoded back to the output feature by 5 layers of 2-D deconvolutional.

CRN benefits from the feature extraction capability of CNNs and the temporal modeling capability of RNNs. Moreover, the skip layer connection is applied to connect each encoder layer to the corresponding decoder layer, which can improve the flow of gradients when the network is deep.

For VAD part, we use a ResCNN, which contains 4 convolution layers, 4 residual blocks (ResBlocks) and 1 fully connected layer (FC) to generate the label of each frame. Fig. 1

TABLE II
AUC(%) COMPARISON AMONG RESCNN BASED APPROACHES WITH NON-JOINT TRAINING APPROACHES(NON-JT) AND JOINT TRAINING APPROACHES(JT) ON NOISE MATCHED AND UNMATCHED CONDITIONS. BOLD FONT INDICATES THE BEST PERFORMANCE.

	noise matched conditions				noise unmatched conditions			
SNR	-5bB	0dB	5dB	Avg.	-5bB	0dB	5dB	Avg.
Non-JT-noi	93.00	94.76	95.44	94.40	80.52	81.29	81.70	81.17
Non-JT-est	93.04	94.86	95.39	94.43	76.06	79.75	84.83	80.21
Non-JT-mid	92.74	94.83	95.38	94.32	75.64	83.78	90.52	83.31
JT-est[9,10]	**93.34**	94.90	95.42	**94.55**	83.44	88.31	91.86	87.87
JT-mid[10]	92.46	94.51	95.15	94.04	82.27	86.89	91.69	86.95
JT-est&noi	93.28	94.85	95.44	94.52	79.90	82.51	85.29	82.57
JT-mid&est[10]	93.08	94.79	95.31	94.39	86.29	89.92	**92.31**	89.51
JT-mid&noi(Pro.)	93.06	**94.91**	**95.54**	94.50	**89.20**	**91.43**	91.82	**90.82**

depicts the ResCNN architecture. The convolutional layer and the residual block are alternately arranged to process the input magnitude spectrum, which can significantly map features into a more separable space. Subsequently, a high-level features learned by these combination blocks are then fed into a fully connected layer with size-1 cell to predict the target.

A more detailed description of the network architecture is provided in Tab. I. The input size and output size of each layer are specified in $timeSteps \times featureMaps \times frequencyChannels$ format. The layer hyper-parameters are given in $(kernelSize, strides, outChannels)$ format. We apply zero-padding to keep the size of time direction. In speech enhancement part, the kernel size is 1×3 $(Time \times Frequency)$, the stride length is 1×2 $(Time, Frequency)$. Note that the number of feature maps in each decoder layer is doubled by the skip connection. In VAD part, the kernel size is 5×5 and 3×3 of convolutional layers and residual blocks, respectively, which can significantly adjust the receptive fields.

III. EXPERIMENTS

A. Experimental Details

All experiments are conducted on TIMIT database [19]. We randomly selected 2000 clean utterances from training set, and use the TIMIT core test set as our test utterances. The TIMIT core test set contains 192 utterances, 8 from each of 24 speakers. We concatenate the selected train utterances with some silence segments of random length, which makes the ratio of speech frames account for around 60%. Then mixed with a speech shape noise (SSN) and 4 other types of noise from the NOISEX-92 dataset [20]: babble noise, factory noise, destroy engine noise, and destroyer operations room noise at SNRs of -5, 0, 5 dB for training. Each noise is divided into two non-overlapping segments for training and testing respectively. To make the sample more generally and multiply, we intercept noise segments from long noise randomly. Besides these four types of noise, another four types of noise are used for noise unmatched test, which includes an unseen factory noise, buccaneer noise from NOISEX-92 and bus noise, street noise from CHiME-4 dataset [21]. All signal is resampled to 16 kHz before mixing.

For speech enhancement, we use noisy and clean speech magnitude spectrum as input and output feature, which utilizes the short-time Fourier transform (STFT) after divided the speech signal into frames using 20 ms hamming window with 10 ms overlap. So the input and output of each frame is 161-dimensional. A log operation is applied to compress the dynamic range and facilitate training. The number of time frames T in the magnitude spectrum is 100 in our experiments.

The input of VAD part is a combination of the original noisy speech magnitude spectrum and the output of the second LSTM layer in speech enhancement part. Finally, we got a 225-dimensional feature map for each frame. We applied the Sohn's method to the clean speech corpus to get the label [22] for VAD. This method was proved to be sufficiently reasonable to generate labels [2], and is same to our comparison approaches [9], [10].

All models are optimized using Adabound optimizer with a mini-batch size of 64, which employ dynamic bounds on learning rates to achieve a gradual and smooth transition from adaptive methods to SGD [23]. We use a constant dropout rate of 0.4 at LSTM layers.

In order to evaluate the performance of the class imbalance problem like VAD, we use the area under the curve (AUC) as the evaluation metrics, which is the area under the receiver operating characteristic (ROC) curve [24]. AUC is considered as an overall metric of the VAD performance rather than the detection accuracy [2], [8]–[10]. Higher value means better performance.

B. Experimental Results

Tab. II lists the comparison results between non-joint-training based approaches and joint-training based approaches under noise matched and unmatched conditions with different SNRs. Non-joint-training based approaches means only the VAD module is applied or the two modules are trained separately. For joint-training based approaches, the hyper-parameter α in Eq. (3) is set to 0.1.

There are three types of input features for ResCNN-based VAD, they are spectrum of noisy speech and estimated speech of CRN-based speech enhancement module which denotes as noi and est, respectively. And the output from the encoder

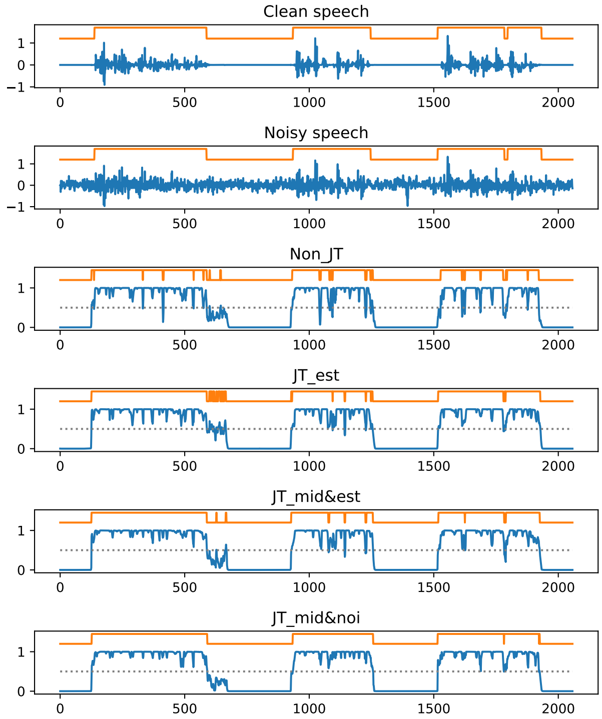

Fig. 2. Visualization the output of an example for the posterior probability (blue) and final decision (orange) of different approaches in babble noise environment at 0 dB SNR level,.

(the second LSTM layer of speech enhancement model, as illustrated in Fig. 1), which denotes as mid. In [9], the input of VAD module is est, exclusively. In [10], the input of VAD module contains est and mid. However, the performance of these approaches are highly dependent on the performance of speech enhancement part, which is more difficult than VAD obviously. So we compared the impact of mid and it's combination. The values of each approach indicate the best results use the same testing set under the same conditions.

By comparing $Non\text{-}JT\text{-}noi$ and $Non\text{-}JT\text{-}est$, we observe that applying speech enhancement first is beneficial for VAD task under noise matched conditions but decrease generalization ability obviously in low SNR. The performance of these methods shows a slightly difference under matched conditions. However, under noise unmatched conditions, the joint training based methods shows better generalization performance, which is benefits from the robust features provided by speech enhancement. More specifically, $JT\text{-}est$ provides over 9.54% relative improvement than $Non\text{-}JT\text{-}est$, and $JT\text{-}mid$ provides over 4.36% relative improvement than $Non\text{-}JT\text{-}mid$ baseline under unmatched condition on average. These results indicate that joint training can improve the performance.

For these three types of input features, we make three types of combination. Where $JT\text{-}mid\&est$ is same to the best model of [10]. Compared with $JT\text{-}mid\&est$, the performance of the proposed $JT\text{-}mid\&noi$ method provides 1.46 % relative improvement under unmatched conditions on average. In low SNR, specifically, the $JT\text{-}mid\&est$ is worse than the proposed $JT\text{-}mid\&noi$. This is mainly because the quality of the estimated magnitude spectrum it depends on has lost some useful information.

TABLE III
AUC (%) COMPARISON WITH AND WITHOUT GRADIENT BALANCE FOR JT-MID&NOI ON AVERAGE. BOLD FONT INDICATES THE BEST PERFORMANCE.

SNR	-5bB	0dB	5dB	Avg.
no GB α=0.05	89.54	92.08	92.99	91.53
no GB α=0.1	91.35	93.36	**93.89**	92.86
no GB α=0.2	88.81	90.88	91.76	90.49
no GB α=0.5	88.75	89.96	90.78	89.83
no GB α=0.8	89.63	92.29	92.72	91.55
with GB	**91.58**	**93.48**	93.74	**92.93**

Fig. 2 shows a visualized comparison performance of the VAD approaches on TIMIT core test set. The first two subgraphs show speech and noisy speech with ground truth, and the rests show the posterior probability and final decision of each method. From the figure and tables, we observe that the output of non-joint training method is noisy compared to joint training methods. Again, the probability curve of the proposed $JT\text{-}mid\&noi$ is smoother and providing the best performance.

Tab. III shows the impact of the hyper-parameter α and the effects of gradient balance. We compare the performance of $JT\text{-}mid\&noi$ approach when α is fixed to 0.05, 0.1, 0.2, 0.5 and 0.8 and α is dynamic adjusted by gradient balance strategy. To apply the gradient balance, initial value of α_0 is set to 0.5, which is not optimal for this task but get similar results. This result indicate that gradient balance can tuning the hyper-parameter automatically.

IV. Conclusions

In this work, we improve the performance of VAD under noise unmatched conditions in low SNR by a joint training method. A CRN-based speech enhancement is employed to get robust features for ResCNN-based VAD. More over, we optimized the speech enhancement and VAD jointly by a dynamic gradient balance strategy. We show that our proposed method has obvious advantages in generalization ability than compared approaches.

V. Acknowledgments

This research was supported in part by the China National Nature Science Foundation (No. 61876214, No. 61866030).

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, April 2013.

[2] X. L. Zhang and D. L. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, Feb 2016.

[3] S. 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5549–5553.

[4] J. Kim, H. Choi, J.Park, J. Kim, and M. Hahn, "Voice activity detection based on multi-dilated convolutional neural network," in *Proceedings of the 2018 2Nd International Conference on Mechatronics Systems and Control Engineering*, no. 5, 2018, pp. 98–102.

[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(ICASSP)*, 2013, pp. 483–487.

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

[7] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," in *Interspeech*, 2016, pp. 3668–3672.

[8] Y. Zhuang, S. Tong, M. Yin, Y. Qian, and K. Yu, "Multi-task joint-learning for robust voice activity detection," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Oct 2016, pp. 1–5.

[9] Q. Wang, J. Du, X. Bao, Z.-R. Wang, L.-R. Dai, and C.-H. Lee, "A universal vad based on jointly trained deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2282–2286.

[10] Y. Jung, Y. Kim, Y. Choi, and H. Kim, "Joint learning using denoising variational autoencoders for voice activity detection," *Proc. Interspeech 2018*, pp. 1210–1214, 2018.

[11] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proceedings of Interspeech*, 2018, pp. 3229–3233.

[12] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, Jan 2014.

[13] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, June 2015.

[14] H. Zhang, X. Zhang, and G. Gao, "Training supervised speech enhancement system to improve stoi and pesq directly," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5374–5378.

[15] L. Sun, J. Du, L. Dai, and C. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, March 2017, pp. 136–140.

[16] T. Gao, J. Du, L. Dai, and C. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5054–5058.

[17] 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)*, May 2014, pp. 2519–2523.

[18] K. He, X. Zhang, S. Ren, and S. Jian, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016,.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, 1993.

[20] A. Varga, Steeneken, and J. Herman, "Assessment for automatic speech recognition ii: Noisex-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535 – 557, 2017.

[22] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[23] L. Luo, Y. Xiong, Y. Liu, and X. Sun, "Adaptive gradient methods with dynamic bound of learning rate," *arXiv preprint arXiv:1902.09843*, 2019.

[24] J.A. Hanley, and B.J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve,"*Radiology*, vol. 143, pp. 29 – 36, 1982.