Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

A Gated Recurrent Unit Based Robust Voice Activity Detector

Il Han
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email:hi_iit@163.com
Chol Nam Om
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: ocniit@163.com
Un Il Kim
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: kui_iit@163.com
Jang Su Kim
Institute of Information Technology, Hightech Research & Development Center
Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea
Email: kjs1982103@163.com

----------**ABSTRACT**----------
Voice activity detection (VAD), which identifies speech and non-speech durations in speech signals, is a challenging task under noisy environment for various speech applications. In this paper, we propose a Gated Recurrent Unit (GRU) based VAD using MFCCs augmented delta and delta-delta features under the low signal-to-noise ratios (SNRs) environments to overcome the shortages of the traditional VAD models. We compare the proposed method with the traditional methods by using speech signals smeared with 10 types of noise at low SNRs. Experimental results reveal that the proposed method based on GRU is superior to traditional method under all the considered noisy environments, indicating that the network based on GRU improve the performance of speech detection.

Keywords -**voice activity detection, deep neural network, recurrent neural network, gated recurrent unit.**

Date of Submission : July 24, 2023 Date of Acceptance: August 19,2023
----

1. INTRODUCTION

Voice activity detection (VAD), which identifies speech and non-speech periods, is an important front-end step for various speech applications, including speech coding, enhancement, recognition and smart elevator and so on. For example, in speech enhancement and in spectral subtraction, speech/non–speech detection, which is applied to detect the signal periods that contain only noise, is used in the noise reduction process [1]. In the digital cellular telecommunication systems VAD is applied to detect non-speech frames, thus reduce average bit rates [2]. VAD is also a very useful technique for improving the performance of speech recognition systems [3].

In prior studies, VAD which take simple acoustic features such as energy and zero crossing rates for detecting speech periods, is suitable for clean signals, but its performance is degraded under noisy environments [4]. However, in various real-life applications, speech signals are always corrupted by the background noises, which cause those simple VAD algorithms to degrade dramatically.

For VAD, a large amount of research under strong noisy conditions has been done [5]. All these methods utilize input such as spectrum-based feature, cepstrum-based features, fundamental frequency-based feature, entropy and harmonic and energy-based features. For example, the long-term spectral divergence between voice and noise were employed in [6], [7]. A VAD algorithm proposed in [8] measures the periodic to aperiodic component ratios to detect speech and non-speech period. In [9], a method based on a Gaussian statistical model is proposed by Sohn et al, where the decision rule is derived from the mean of the likelihood ratio for individual frequency bands by assuming that the noise is already known. The drawback of this methods performs well under stationary noise, but their performance is degraded under non-stationary noise.

The performance of VAD using machine learning methods is superior to previous methods. For example, SVM methods **Error! Reference source not found.** and deep neural networks (DNNs) based methods **Error! Reference source not found.**, **Error! Reference source not found.** have been found that its performance is better with traditional VAD.

Recently, with the advent of artificial neural networks (ANNs) in the form of deep learning algorithms, neural network-based VAD has become very popular. In [15],

Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

Espi et al proposed convolutional neural network (CNN), which utilized spectro-temporal features as the input, to detect non-speech acoustic signals. A deep maxout DNN was proposed to improve the VAD performance [16]. Zhang et al [17] utilized the combination of multiple features, such as MFCCs, pitch, DFT and so on, as the input to DNN to optimize the capability of DNN-based VAD.

With the development of deep learning, especially CNN [18], [19] and recurrent neural networks (RNN)[20]-[22] based VAD have used in successful applications. Recent works in VAD have focused to improve the robustness towards noise where the training data have smeared with noise by corrupting clean speech with foreground or background noise.

The prior VAD algorithms are based on the assumption of quasi-stationary noise, which means the noisy signal changes much slower than the speech signal. Also these algorithm makes the decisions from the current. The statistical VAD are based on the assumption that the frequency bins in one frame are statistically independent. However, there are noise signals such as clapping and clanks that change faster than the speech signal. The frequency bins in the same frame can be utilized by processing them together from highly correlation of the consecutive audio frames. Still, this assumptions mentioned above, which work well in all speech processing system.

However, the main drawback of DNN for detecting speech is that they ignore the local temporal and spectral correlation in the speech features. The drawback of CNN in modeling speech signals is that it doesn't consider long term temporal dependencies. Thus, DNN and CNN are not suitable for time series signal processing task. In many sequence modeling tasks, especially speech recognition, machine translation and language modeling [23], RNNs have shown superior performance. The reason is that RNNs not only utilize the temporal relation between the input signals, but also consider the long-term dependencies.

GRU (Gate Recurrent Unit) is a variant of RNN and better than LSTM. Some experiment results on small datasets show that the GRU are faster to train and less to diverge than LSTM. Motivating this, we proposed a deep neural network for VAD based on GRU.

The main contributions of our paper are as follows. First, we propose a deep neural network for VAD. The extracted MFCC features augmented delta and delta-delta from past frames are fed to neural network and give decision whether current frame is speech or not. We construct network with TDNN and GRU with good advantage for exploring the temporal relation and long-term dependency effectively. Second, we trained this network in a supervised method and evaluated in various noises under low SNRs environments with other methods. The rest of this paper is organized as follows. In Section 2, we describe the proposed neural networks for VAD using speech dynamics. In Section 3, we provide the experimental results and discussion of the results. Finally, we give the conclusions.

2. PROPOSED METHOD

2.1 FEATURE EXTRACTION

Every utterance has a continuous speech duration which has a start and an end. For VAD, it is important to detect a start point and end point of speech signal. It is simple for clean speech (SNR is greater than 30dB) but the performance is degraded under noisy environment. We utilize delta and delta-delta of log energy of the frame for detecting the candidate of start point and end point.

For VAD, it is essential to extract speech feature from speech signal. Experimental Result shows that speech dynamics, such as delta and delta-delta cepstrum, are more effective for modeling speech. MFCC (Mel-frequency cepstral coefficients) is an acoustic feature, which mimics well the production and reception system of human speech. The human ear receives frequencies less than 1 KHz at linear scale, but receives frequencies higher than 1 KHz at logarithmic scale. MFCCs uses the property of human ear, thus we take this as acoustic feature. Equation (1) shows Mel scale.

$$m(f) = 1125 \times \ln(1 + f/700) \quad (1)$$

Adding delta and delta-delta features, which are computed as the 1st and 2nd order derivation of MFCCs, the better result can be obtained. We didn't consider the 3rd derivation of MFCCs since it has no improvement of speech recognition. That is why delta and delta-delta cepstra are features that express dynamics referring to the time-varying properties of speech signals [24]. In result we consider 39 dimensional features, 13 MFCCs augmented with delta and delta-delta features. The sampling rate of signal is 16 KHz.

The input signals are divided into frames whose length is 20ms. The segment consisting of past frames is framed into overlapping frames of length l (20ms) with a stride (10ms) by using Hamming window, giving a total of $N=(L-l)/s+1$ frames. From each frame, 39-dimentional MFCCs augmented delta and delta-delta features are extracted, we generate a total of $N \times F$ features for the input signal segment of length L. In this way we obtain a set of frequency-domain spectral coefficients from time-domain speech signals.

After getting the MFCC matrix in segments, the extracted features which are fed to neural network, are normalized to zero mean and unit standard deviation. It makes decision for the current frame i.e. voice and non-voice.

Because the VAD is a binary classifier, it outputs for each frame a binary vector whose elements are determined as 1 or 0.

Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

After performing hang-over scheme, we get final VAD output series consisting of 1s (ones) representing the speech and 0s representing the non-speech.

2.2 GRU-BASED VAD

Fig. 1, 2 and 3 shows the proposed method, the structure of GRU unit and the architecture of the proposed neural network, respectively.

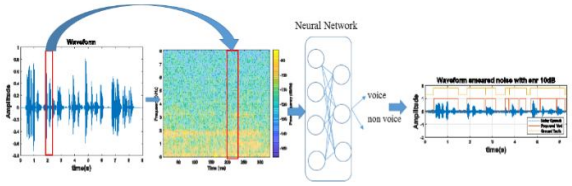

Fig. 1. Proposed method pipeline

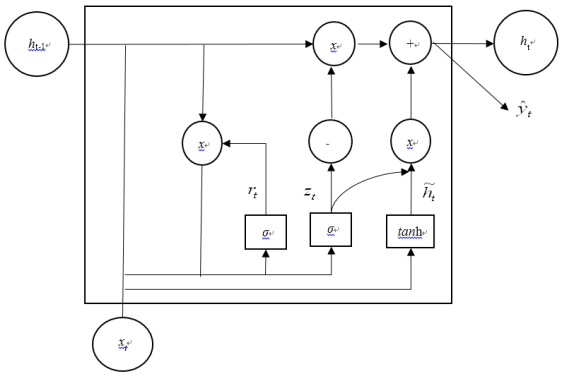

Fig.2. Structure of GRUunit

DNN in [24] is composed of five layers of RBMs. Out method use two RNN layers instead of 2 RBM layers.

The Gated Recurrent Unit (GRU), which is widely used in many areas, is another variant of traditional RNN [25]. GRU, which uses fewer parameters than LSTMs [26], [27], is used as the base cell for recurrent layers. Since the input gate and the forget gate in LSTM are combined into one gate operation, it has a simpler structure of gate operations than that of LSTM. One hidden node of GRU is consisted of the candidate activation g and two kinds of gate operations such as the reset gate r and the update gate z.

It consists of 5 time-delayed deep neural network (TDNN) layers and 2 stacked GRU layers. Three TDNN layers is at the bottom, which are followed by GRU [25], [28]. The DNN layer has 256 units and the activation function of each DNN is Rectified Linear Unit (ReLU). The feature matrix from sequence are fed to first DNN layer as input. The output followed by ReLU is inputted to second DNN layer. The output of second layer is fed to third DNN layer. The output of third DNN are fed to RNN layer, where RNN cell is GRU. The unit of this layer is 256. The output of RNN is fed to fully-connected layer, which give probability of speech. Here, the recurrent layer of RNN is bi-directional [29], [30].

The context specification of each TDNN layer is as follows. The input x_1^t at time t of 1st TDNN layer is determined from signal $\{x_0^{t-2}, x_0^{t-1}, x_0^t, x_0^{t+1}, x_0^{t+2}\}$. The input x_2^t at time t of 2nd TDNN layer is determined from signal $\{x_1^{t-1}, x_1^t, x_1^{t+1}, x_1^{t+2}\}$. The input x_3^t at time t of 3rd TDNN layer is determined from signal $\{x_2^{t-3}, x_2^t, x_2^{t+3}, x_2^{t+6}\}$.

The output of 3rd TDNN is fed to two stacked GRU layers. The output of GRU layers is fed fully-connected layer and give VAD decision of the frame finally.

Since the weights of RNNs are reused across all the time steps, the RNNs have less number of parameters than DNNs. We applied layer normalization **Error! Reference source not found.** to be beneficial for training GRUs. During each time step the hidden states of GRU layers are normalized.

After determining the speech and non-speech of frames, we performed hangover method. We determined as silence segment if the number of consecutive frames labeled 0 is greater than 15(0.3seconds length). Else we set all these frames labeled 0 as voice frame.

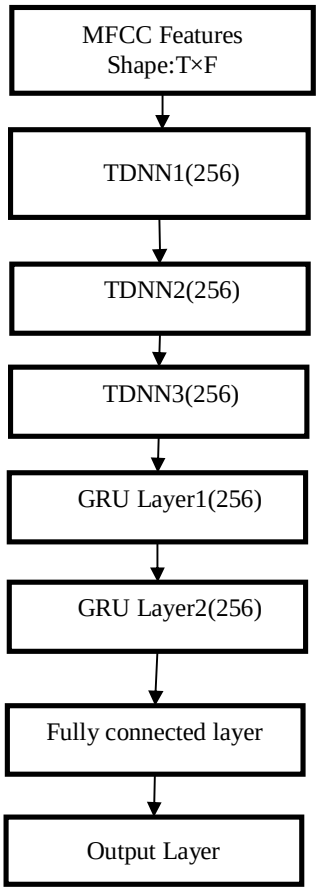

Fig.3. Architecture of Neural Network for VAD

Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 DATASET AND EXPERIMENTAL SETUP

In the experiments, we use speech files from TIMIT [32]. In TIMIT corpus, every utterance are clean speech. The size of the dataset is about 12GByte and the size of the records is 1.56*1000000.Single utterance contains a small number of non-speech periods so that it is difficult to evaluate a VAD algorithm accurately. After concatenating a number of utterances which was chosen randomly from every collections of utterances into a single recording, we appended a few seconds of silence at the beginning of utterance. Also silence was appended at the ending and junctions of the utterances. To equalize the power we normalized the amplitudes of each utterance.

Let $a \in \mathbf{R}$ be the audio signals. Then we obtained the initial labels (if speech is 1else 0) of every frames of utterance by VAD which utilized simple energy. Our proposed method performs a frame by frame classification for VAD, we only consider T past frames. We construct from audio signal a dataset of overlapping frames by concatenating consecutive frames into sequence of length$T$. We labeled each sequence according to the label (1 or 0) of the last frame of the sequence because only past frames are used for classification. That is, the sequences S_a^i containing audio frames are labeled to the label of the *i-th* frame of the signal.

Also, we used a collection of a large number of different noise signals from [33]. To obtain noisy signals, the clean speech files were mixed with ten types of noises from NOISEX-92 [30]. Each segment consisted of 30 frames, which means input vector has a hundred of elements. Each noise signal is differently selected for the speech files with SNRs at 10, 5, 0, and −5dB.

The noise types used in this evaluation are as follows: factory 1, leopard, m109, buccaneer1, buccaneer2, babble, machinegun, hfchannel, white and pink. Others are non-stationary except pink and white noise. We smeared noises to speeches at a desired SNR using MATLAB. Thus, we create a dataset of clean/noisy pairs for training.

We divided these pairs 80%/10%/10% into training/dev/test set. During training the parameters of model was estimated in development sets and the performance was evaluated in test sets.

Our network was trained in TensorFlow Framework using the weighted cross-entropy and Adam optimizer [34]. All weights of network were initialized with values from a Xavier initializing scheme. The model is trained for 10K iterations with a batch size of 100. The initial learning rate is 5×10^4 and reduced 5 times after each 2K iterations. We use dropout with the probability of 0.5 on the output. We also use batch normalization on audio output. After training, the performance of the model is evaluated on the test set.

3.2 RESULTS AND DISCUSSION

To compare the performance of our method, some traditional VAD methods, such as LTSD [5], LTPD [6], LSFM [7], Sohn [9] and DNN-based VAD [24] which have been known as noise robust method, are taken for comparison. Also we compared the performance of these neural networks for VAD under ten noises with SNRs at 10dB, 5dB, 0dB and -5dB.

To compare the performance of the different methods, the receiver operation characteristic (ROC) curve, which represents the classification capability, is taken. The TPR and FPR are defined as follows:

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN} \quad (2)$$

, where TP means true positive, FP means false positives, FN means false negatives, and TN means true negatives.

However, TPR and FPR are multiple number evaluation metric. To get the quantitate value of ROC of various VAD methods, the area under the curves (AUCs), which are the main single-number evaluation metric, are calculated. Table 1 shows the average AUCs under ten type's noises under -5dB SNR of various methods and neural network based VAD algorithm.

Table. 1. AUC comparisons of the evaluated algorithms under -5dB SNR

	Sohn	LTSD	LSFM	LTPD	FNN	Proposed
Factory1	0.5538	0.5978	0.7113	0.8998	0.8261	0.9286
Leopard	0.9608	0.8721	0.9623	0.9435	0.9883	0.9999
m109	0.9182	0.8498	0.8561	0.8696	0.6597	0.8331
buccaneer1	0.7612	0.8471	0.7921	0.9382	0.9389	0.9441
buccaneer2	0.8162	0.8794	0.9086	0.9495	0.9022	0.9526
Babble	0.7687	0.8556	0.6873	0.7788	0.6712	0.8629
hfchannel	0.8814	0.9134	0.8626	0.9312	0.9049	0.9283
car	0.5934	0.7860	0.3423	0.9380	0.9611	0.9944
Pink	0.7802	0.8609	0.8777	0.9481	0.8902	0.9662
White	0.8601	0.8901	0.9096	0.9521	0.9853	0.9895
Average	0.8172	0.8532	0.8352	0.9213	0.9228	0.9401

As shown in Table 1, GRU based VAD is better than traditional and DNN based VAD. Especially the AUC of GRU based VAD improves by 2~15% than Sohn, LTSD, LSFM and LTPD. Also it improves by 7~9% than DNN.

Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

To evaluate the performance, we also measured the TPR (sensitivity) and TNR (specificity). In VAD problem, Sensitivity (Speech hit rate) means the percentage of the frames correctly classified as speech among all the speech frames, and specificity (Non-speech hit rate) means the percentage of the frames correctly classified as non-speech among all the non-speech frames [35].

Table 2 shows the mean TPR and TNR for SNR levels range from -5dB to 10dB. As shown in the table, the GRU based method has a high sensitivity and specificity than traditional methods and DNN. Interestingly, the performance of the GRU-based VAD method outperforms the others in finding speech, particularly for low SNRs and non-stationary cases.

Table. 2. Average speech/non-speech hit rates for SNR levels range from -5dB to 5dB

	Sohn	LTSD	LSFM	LTPD	FNN	GRU-RNN
TPR(HR1) (%)	94.25	98.28	87.77	94.23	90.47	94.96
TNR(HR0)(%)	59.80	38.92	76.00	87.77	89.45	92.30

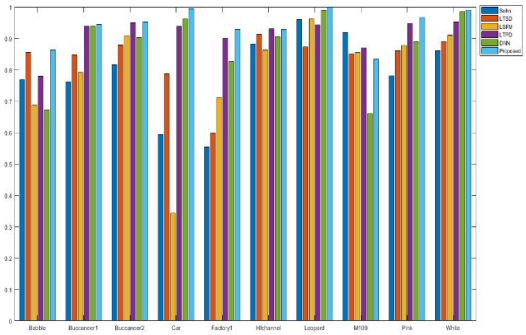

Fig. 4. Comparison of AUC with different methods under -5dB SNR

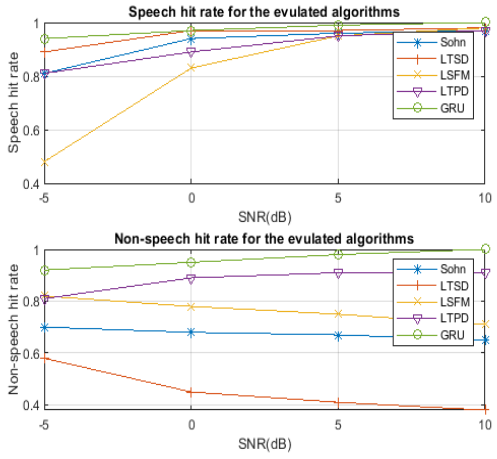

Fig. 5. Speech/non-speech hit rates of the evaluated algorithms under different SNR levels

In Fig. 4, we show the comparison result of different methods. Fig. 5 shows the speech/non-speech hit rates of the proposed method based GRU and traditional VAD methods respectively. As shown in the above figures, the GRU-based method shows an advantage over the other VAD methods. The GRU based VAD is more effective for low SNR cases. In the cases of stationary noise, such as white and pink noise, these methods achieve a high TPR and a low FPR. In the cases of non-stationary noise, such as babble and m109 noise, the AUC of the GRU-based method is higher than others by 9~11%.

Totally, the performance of the GRU based VAD is superior to that of the traditional methods and DNN based VAD method mainly due to considering the relation of consecutive frames and bins.

4. CONCLUSIONS

In this paper, a VAD algorithm based on GRU neural network for improving the performance of VAD was presented, We determine whether current frame is speech or non-speech frame from the past frames. The extracted MFCCs with delta and delta-delta features of segment are fed to deep network. After performing hang-over method, we obtain the final VAD. To evaluate the performance of the proposed method, speech signals smeared with ten kinds of noise, such as white, babble, factory, car, pink and so on were tested at SNRs of 10, 5, 0, and -5dB.

The experimental results show that the neural network based on GRU, reflecting the time series characteristics of the speech signal, is more effective than other methods for VAD under considered noisy conditions.

REFERENCES

[1] S.F. Boll, Suppression of Acoustic Noise in Speech Using Spectral Subtraction, IEEE Transactions on Acoustics, Speech and Signal Processing, 27(2), 1979, 113-120.

[2] A. Benyassine, E. Shlomot, H.Y. Su, D. Massaloux, C. Lamblin, and J.P. Petit, ITU-T Recommendation G.729 Annex B: a Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications, IEEE Communications Magazine, 35(9), 1997,64-73.

[3] S.B. Tong, N.X. Chen, Y.M. Qian, and K. Yu, Evaluating Vad for Automatic Speech Recognition, Proc. 12th International Conf. on Signal Processing, Hangzhou, PRC, 2014,2308–2314.

[4] L. Rabiner, and M.R. Sambur, An Algorithm for Determining the Endpoints of Isolated Utterances, Bell System Technical Journal, 54(2), 1975,297-315.

[5] J. Ramirez, J.C. Segura, C. Benitez, A.de la Torre, and A. Rubio, Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information, Speech Communication, 42(3-4), 2004,271-287.

[6] X.K. Yang, L.He, D. Qu, and W.Q. Zhang, Voice Activity Detection Algorithm Based on Long-Term

Volume: 15 Issue: 02 Pages: 5831 – 5836 (2023) ISSN: 0975-0290

Pitch Information, EURASIP Journal on Audio, Speech and Music Processing, 2016:14, 2016,1-9.
[7] Y.N. Ma, and A. Nishihara, Efficient Voice Activity Detection Algorithm Using Long-Term Spectral Flatness Measure, EURASIP Journal on Audio, Speech and Music Processing, 2013:21, 2013,1-18.
[8] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazak, Noise Robust Voice Activity Detection Based on Periodic to Aperiodic Component Ratio, Speech Communication, 52(1), 2010,41-60.
[9] J.S. Sohn, N.S. Kim, and W.Y. Sung, A Statistical Model-Based Voice Activity Detection, IEEE Signal Processing Letters, 6(1), 1999,1-3.
[10] E.Q. Dong, G.Z. Liu, Y.T. Zhou, and X.D. Zhang, Applying Support Vector Machines to Voice Activity Detection, Proc. 6th International Conf. on Signal Processing, Beijing, PRC, 2002,1124–1127.
[11] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H.Z. Li,Voice Activity Detection Using MFCC Features and Support Vector Machine,Proc. International Conf. on Speech and Computer, 2007,556–561.
[12] Q.H. Jo, J.H. Chang, J.W. Shin, and N. S. Kim, Statistical Model-Based Voice Activity Detection Using Support Vector Machine, IET Signal Processing, 3(3), 2009,205-210.
[13] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza, A Deep Neural Network Approach for Voice Activity Detection in Multi-Room Domestic Scenarios, Proc. International Joint Conf. on Neural Networks, Killarney, IRELAND, 2015,1–8.
[14] 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, 24(2), 2016,252-264.
[15] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, Exploiting Spectro-Temporal Locality in Deep Learning Based Acoustic Event Detection, EURASIP Journal on Audio Speech and Music Processing, 2015:26, 2015,1-12.
[16] S.M. Valentin, N.P. Tatiana, and A.P. Alexey, Robust Voice Activity Detection with Deep Maxout Neural Networks, Modern Applied Science, 9(8), 2015,153-159.
[17] X.L. Zhang, and J.Wu, Deep Belief Networks Based Voice Activity Detection, IEEE Transactions on Audio, Speech and Language Processing, 21(4), 2013,697-710.
[18] 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,Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing, Calgary, CANADA, 2018,5549–5553.
[19] A. Sehgal, and N. Kehtarnavaz, A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection, IEEE Access, 21, 2018,9017-9026.
[20] M. Lavechin, M.P. Gill, R. Bousbib, H. Bredin, and L.P. Garcia-Perera, End-to-End Domain-Adversarial Voice Activity Detection,Proc. Conference of the International Speech Communication Association, Shanghai, PRC, 2020,3685–3689.
[21] T.J. Xu, H. Zhang, and X.L. Zhang, Polishing the Classical Likelihood Ratio Test by Supervised Learning for Voice Activity Detection,Proc. Conference of the International Speech Communication Association, Shanghai, PRC, 2020,3675–3679.
[22] Z.P. Zheng, J.Z. Wang, N. Cheng, J. Luo, and J. Xiao, MLNET: an Adaptive Multiple Receptive-Field Attention Neural Network for Voice Activity Detection,Proc. Conference of the International Speech Communication Association, Shanghai, PRC, 2020,3695–3699.
[23] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur, Recurrent Neural Network Based Language Model,Proc. Conference of the International Speech Communication Association, Makuhari, JAPAN, 2010,1045–1048.
[24] S. Dwijayanti, K. Yamamori, and M. Miyoshi, Enhancement of Speech Dynamics for Voice Activity Detection using DNN, EURASIP Journal on Audio, Speech and Music Processing, 2018:10, 2018,1-15.
[25] K.H. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation, arXiv preprint, arXiv:1406.1078, 2014.
[26] S. Hochreiter, and J. Schmidhuber, Long Short-Term Memory, Neural computation, 9(8), 1997,1735-1780.
[27] F.A. Gers, N.N. Schraudolph, and J. Schmidhuber, Learning Precise Timing with LSTM Recurrent Networks, Journal of Machine Learning Research, 3(1), 2003,115-143.
[28] J.Y. Chung, C. Gulcehre, K.H. Cho, and Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv preprint, arXiv:1412.3555, 2014.
[29] M. Schuster, and K. K Paliwal, Bidirectional Recurrent Neural Networks, IEEE Transactions on Signal Processing, 45(11), 1997,2673-2681.
[30] Noisex-92 Database, Rice University, Available at: http://spib.linse.ufsc.br/noise.html. Accessed on 22 Feb 2017.
[31] J.L. Ba, J.R. Kiros, and G.E. Hinton, Layer Normalization, arXiv preprint, arXiv:1607.06450, 2016.
[32] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, Darpa Timit Acoustic-Phonetic Continuous Speech Corpus CD-ROM, NIST Interagency/Internal Report, NISTIR-4930, NIST, Gaithersburg, 1993.
[33] 100 Nonspeech Environmental Sounds, Available at: http://www.pudn.com/Download/item/id/3457634.html ,2018.
[34] D. Kingma, and J. Ba, Adam: a Method for Stochastic Optimization, arXiv preprint, arXiv:1412.6980, 2014.
[35]R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, 2nd edn, Wiley-Interscience, New York, 2001.