Robust Voice Activity Detection Based on LSTM Recurrent Neural Networks and Modulation Spectrum

Phuttapong Sertsi, Surasak Boonkla, Vataya Chunwijitra, Nattapong Kurpukdee, and Chai Wutiwiwatchai
National Electronics and Computer Technology Center (NECTEC),
National Science and Technology Development Agency (NSTDA),
112 Pahonyothin Road, Pathumthani, 12120, Thailand
Email: {phuttapong.sertsi, surasak.boonkla, vataya.chunwijitra, nattapong.kurpukdee, chai.wutiwiwatchai}@nectec.or.th

***Abstract*—Voice activity detection (VAD) used for classifying speech/non-speech sections of a speech signal still suffers from noisy environments. In this paper, we cooperate the modulation spectrum (MS) and the long short-term memory recurrent neural network (LSTM) to improve the robustness. The baseline LSTM used conventional speech features in training and classifying speech and non-speech sections. The proposed VAD system by using MS as another speech feature in order to increase the robustness. In addition, we propose a new approach for the computation of the MS feature. The results showed that the accuracy of using the proposed method can be improved under both seen and unseen noise conditions compared with the baseline technique.**

I. Introduction

Voice activity detection (VAD) is an important method to distinguish speech/non-speech sections of a speech signal. It is widely applied in many applications such as automatic speech recognition (ASR) systems, speaker recognition, audio conference system, hands-free telephony, etc. Since, real environments consist of various noises, the VAD must be able to precisely classify speech/non-speech under noisy conditions. The performance of the conventional VAD is shown in Fig. 1 where the red dashed lines indicate the speech/non-speech segments. In Fig. 1(a), most of the VAD system can classify speech/non-speech sections under clean condition. However, the performance of the VAD systems will decrease when various noises exist such as musical and environmental noises as shown in Fig. 1(b).

In the last decade, several VAD methods have been proposed. They are divided into two groups. The first group exploits speech features such as energy thresholds and pitch [1], zero crossing rate [2], fundamental frequency and cepstral feature [3], and modulation spectrum [4]. The second group is model based methods for example, Gaussian mixture model [5], neuron network, recurrent neuron network [6]. Most of the methods in the first group are robust when the speech features are robust in noisy environments. The existing proposed methods, however, yield low performance when various noises exist. Moreover, using speech features usually requires criteria for thresholding which is not valid in all conditions [7].

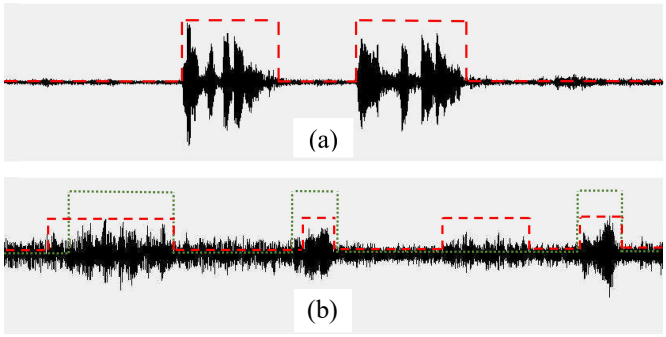


Fig. 1: An example of VAD: (a) a clean speech signal and (b) a noisy speech signal. The green dot line represents the correct speech/non-speech sections of the speech signal and the red lines represent the detected speech/non-speech sections by using VAD.

To avoid such thresholding, the model based VAD methods are employed. It also allows the combination of multiple speech features into a vector without worrying about the criteria for speech/non-speech classification. To date, long short-term memory (LSTM) has become state-of-the-art in the current model based systems [8]. Although it is robust in noisy conditions, its performance still drops when the characteristics of noises are similar to those of speech signals, such as musical and environmental noises in low signal-to-noise ratio (SNR) conditions.

Therefore, we propose a robust VAD based on LSTM by combining the modulation spectrum (MS) in addition to a commonly used speech features: mel-frequency cepstral coefficient (MFCC) and pitch, to improve the robustness of the VAD when various kinds of noises exist. The reason of using LSTM is that it is capable of learning the dynamic of the input and adaptable using previous input for determining the current frame. To improve the robustness, the MS is applied in our system. The difference of MS characteristics between speech and noise is expected to be helpful in

improving the ability of VAD. In addition, we define a new approach to calculate the feature of MS which is flexible than that used before [7]. The remaining of this paper is organized as follows. Section II describes necessary background knowledge. The proposed method is addressed in section III. The experiment and evaluation of this proposed are shown in section IV and V. Finally we discuss and conclude this work in Section VI and VII.

II. Background Knowledge

A. LSTM Recurrent Neural Networks

Neural networks have grown in popularity in the recent years. Nowadays, a speech classification systems are created by using neural networks [6]. One popular method is LSTM, which is able to manage long-range dependencies between the inputs. It is widely used in speech classification, especially VAD system [8]. The main idea of LSTM is a decision gate in cell state. The single cell state is shown in Fig. 2. The gate in LSTM does have the competency to remove or add information to the cell state. Each gate is operated by sigmoid function, which output numbers between zero and one. The output of zero means the data can't send through the gate. On the other hand, the output of one means the data can send through the gate. An LSTM consists of input gate, forget gate and output gate which manage data to enter, delete and leave depending on context and previous output.

An LSTM network calculate N sequence of input $x = [x_1, ..., x_N]$ and output sequence $y = [y_1, ..., y_N]$ by calculating the network to adjust the weights w of each memory cell C_t. While data send through the cell, the data will be calculated by each gate as shown in Fig. 2. When the input x arrived at the memory cell input. Firstly, the input x will be used to compute at the input gate which determines what new information to keep in the memory cell. This decision is made by using sigmoid and hyperbolic tangent function (tanh). Secondly the next step is to decide what information from the previous cell will be thrown away from a memory cell. This procedure is calculated by the forget gate. Thirdly, both data from input gate and forget gate will combine to create an update to the cell state. Finally, to send output to the next cell, the output will be computed by sigmoid and tanh function at the output gate.

B. Modulation Spectrum

Fourier transform of a sufficiently short segment of speech signal yields a short-term spectrum of speech. This short-term spectrum is a vector describing 10 - 30 ms of the speech signal. This short-term analysis plays an important role in speech processing for more than half-century. The spectrogram is a series of there short-term spectra as shown in Fig. 3(a) where the vertical axis is the frequency and the horizontal axis is time. The spectrogram a speech signal describes frequency content of that speech which varies over time.

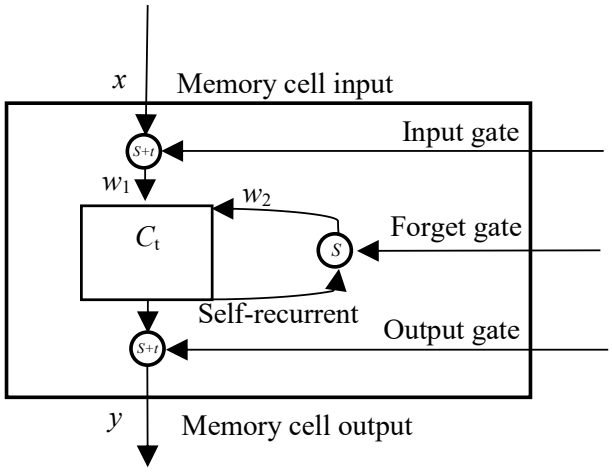


Fig. 2: Illustration of an LSTM memory cell.

On the other hand, the modulation spectrum is the temporal fluctuation of the speech signal [8]. In Fig.3(a), each short-term spectral vector has K number of elements each of which is associated with the k-th frequency bin. If we consider the k-th frequency bin as shown in the red box of this figure within 1000 ms, we will get the temporal fluctuation as shown in Fig. 3(b). The Fourier transform of this fluctuation gives the MS as shown in Fig. 3(c) which has specific features such as the location of the peak, bandwidth or Q-value, and the slop after the peak [7].

The MS of noises will have peaks around 2 Hz to 5 Hz of the modulation frequency as shown in Fig. 3(c). This figure shows the MS of speech signals, environmental noise, and a stationary noise signal. Usually, the peak of noise will appear at low modulation frequency whereas the peak of the speech signal is around 4 - 16 Hz. In addition, the Q-value and the slope of MS after the peak are different as well [7]. Therefore, these features of MS can be used for the VAD.

III. LSTM-Based VAD

The LSTM-based VAD is generally trained by using noisy speech signals to improve its robustness in noisy environments. Sometimes this system is referred to as noise assisted (NAT) VAD system [5]. The robustness comes from the robust input speech features. The traditional VAD based on LSTM is usually trained by the features MFCC, energy, and pitch which is defined as

$$o_t = \{x_1, x_2, x_3, ..., x_{39}, p_1, p_2, p_3\} \quad (1)$$

the x_1 to x_13 are 12 MFCC coefficients and energy. x_14 to x_39 are the first and second derivatives of x_1 to x_13. $p1$, $p2$, and $p3$ are the information of pitch [10]. In this paper, we propose to append the information of MS to the above feature vector so that the new feature is written as

$$s_t = \{x_1, x_2, x_3, ..., x_{39}, p_1, p_2, p_3, ms_1, ms_2, ms_3\} \quad (2)$$

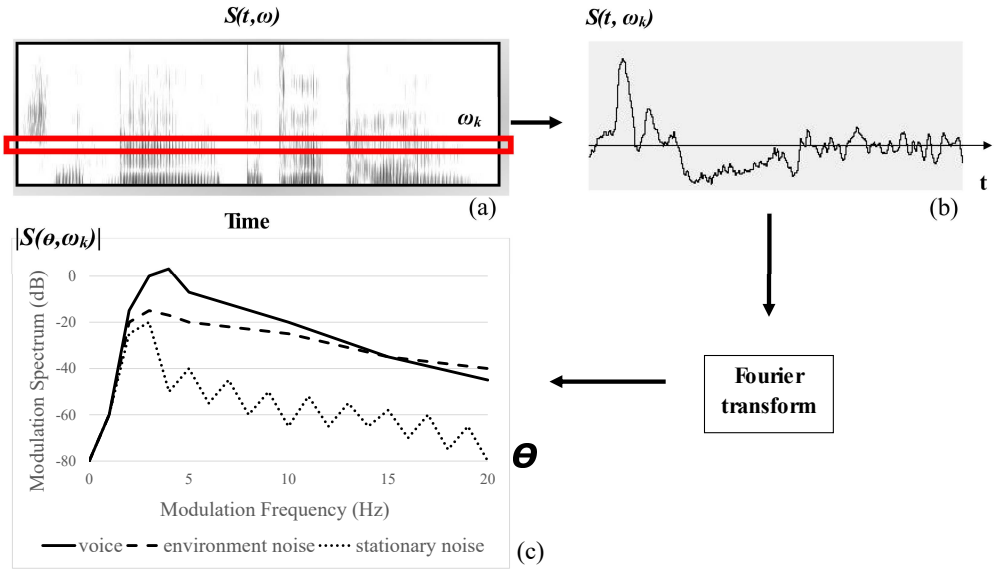


Fig. 3: Modulation spectrum of speech.

where, ms_1, ms_2, ms_3 are the feature of MS, its first, and second derivatives, respectively. There are three parameters to represent the features of MS which are the location of the peak, Q-value, and its slope after the peak of MS [7]. In this paper, we choose the Q-value to represent the feature of MS of speech and noise. The Q-value is calculated from the smooth pattern of MS obtained from several frames of speech signals. However, the insufficient number of frames results in the variation of MS leading to the error in the Q-value calculation. Therefore we use the area under the curve of MS which is calculated by

$$ms_1 = \int_0^{f_c} \log |S(\theta, \omega_k)| d\theta \tag{3}$$

where θ is the modulation frequency, ω_k is k-th frequency bin and f_c is the cut-off frequency which is 20 Hz.

IV. Experiments and Evaluation

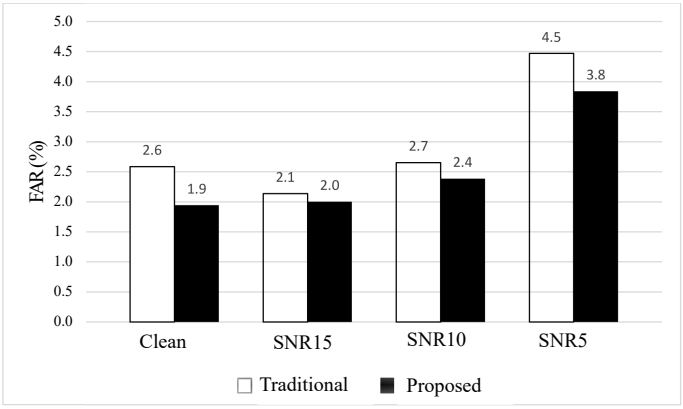
The objective of the experiment is to determine the robustness of the proposed method. It was tested under several kinds of noises under three values of SNRs. The proposed method was trained by using noisy speech signals generated from adding noises to clean speech signals. The clean speech signals were 22 hours of speech data of 40 speakers (20 male and 20 female) from the speech database, namely LOTUS [11]. Noises consisted of airport, babble, car, exhibition, restaurant, street, subway, and train noises from Aurora5 [12], and SNRs were 15, 10, and 5 dB.

The evaluation data consisted of three sets. The first set was clean speech signals from eight speakers which exclude from the training set. Each of them spoke 35 sentences. The second set was for training. It contained the noisy speech signals generated from adding noises to the clean speech signals in the first set with SNRs were 15, 10, and 5 dB. The third set was an hour of 14 sentences from real environments with unseen noise conditions under SNRs ranging from 5-30 dB. The real environments include meeting rooms, streets, and markets.

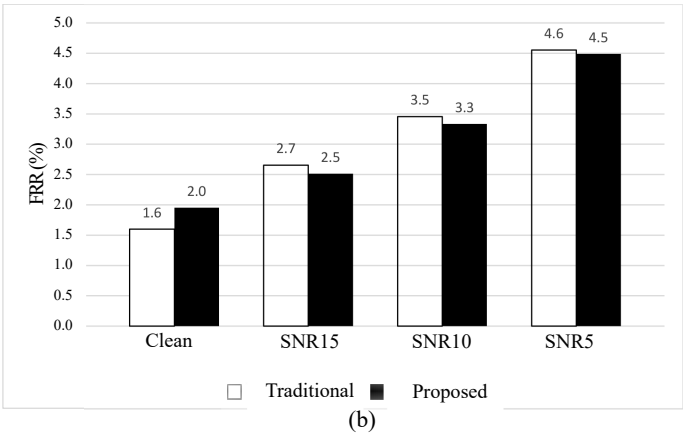
MFCC and pitch were extracted from the training data by using Kaldi toolkit [10]. The MS feature was calculated from the first feature of MFCC. Therefore, a feature vector consisted of 13 mel frequency cepstral coefficients, pitch, and MS including their first and second derivatives. Thus the dimension of the feature vector was 45. The frame length and shift were 25 and 10 ms, respectively.

The training was done by using Keras toolkit [13]. The input layer had 45 input nodes associated with the dimension of the input feature vector. The hidden layer of the network was 256 LSTM layer with hyperbolic tangent activation, densely-connected layer with softmax activation function and the dropout value of which was 0.25%. The backpropagation through time (BPTT) algorithm used 10^{-3} learning rate [14].

In this evaluation, the traditional method was created by using conventional MFCC, energy, and pitch compared with the purposed method which appends the feature of MS after MFCC, energy and pitch. There are two experiments: (I) we used the clean and noisy speech signals in three SNR conditions (+15dB, +10dB and +5dB) to evaluate the robustness of our VAD system. (II) we used unseen noisy speech signals recorded from real environment to test the performance of our VAD system in real-life conditions. The evaluation was done by using false acceptance rate (FAR) and false rejection rate (FRR). FAR is the probability to decide a voiced frame during a silent interval which can be calculated by



(a)



(b)

Fig. 4: (a) FAR and (b) FRR between traditional method and proposed method under noisy conditions

$$FAR = \frac{N_{FA}}{N_S} \times 100(\%) \tag{4}$$

where N_S is a total number of frames and N_{FA} is a number of non speech frames detected as speech frames. Similarly, FRR is the probability to decide a silence frame during a speech interval which can be determined by

$$FRR = \frac{N_{FR}}{N_S} \times 100(\%) \tag{5}$$

where N_{FR} is a number of speech frames detected as non speech frames.

V. Result

A. Experiment I: comparison under clean and noisy conditions

To investigate the accuracy of our VAD system under the clean and noisy conditions, we compare the performance of our proposed VAD system and that of the traditional

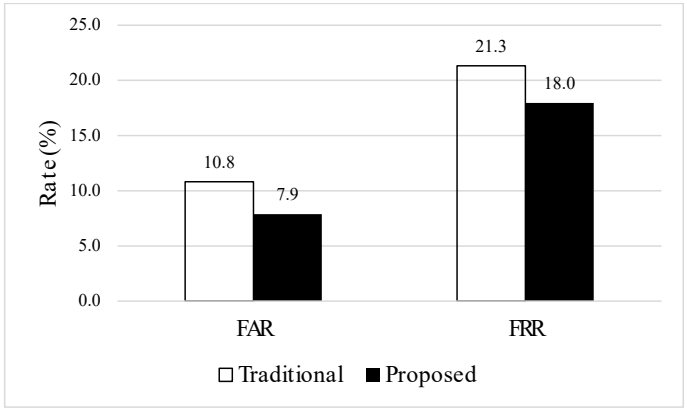


Fig. 5: FAR and FRR between traditional method and proposed method under real environments

baseline LSTM by using the FAR and FRR. The result is shown in Fig. 4 where the white bars and black bars represent the baseline and proposed system respectively. The FARs and FRRs of both methods increase as SNR decreases. However, those of the proposed method are slightly less than those of the baseline. FRR shows the same trend as FAR. These results suggest that our proposed method can determine the speech and non-speech sections under noisy environments.

B. Experiment II: comparison under real environments

To investigate the performance of our proposed approach under a real-life conditions, we evaluate the noisy speech recorded from real environments. The result of the second experiment is shown in Fig. 5, where the white bars represent the traditional baseline and the black bars represent our proposed method respectively. The FARs and FRRs of our method are slightly less than those of the traditional method. These results suggest that our proposed method can classify the speech and non-speech sections under a noisy conditions in real environments.

VI. Discussion

According to the results from two experiments, we can see that our LSTM-based VAD system using MS yields high accuracy as well as the baseline under clean conditions. However, the FRR of the proposed method in the experiment (I) was slightly lower than traditional because the proposed method was trained using noisy speech signals. The clean speech signals which are similar to noise such as fricatives can be classified as noise which causes the FRR lower than that of clean speech. However, our proposed method is better than the baseline in noisy conditions. The FARs and FRRs were slightly less than those of the traditional method which suggests that our method can efficiently classify speech or non-speech under low SNR. In experiment II, the FAR and FRR decreased around 3.5% and the

performance improvements of our proposed technique are statistically significant at the 0.05 level for FAR and FRR when compared with the traditional method. This result justifies that our proposed VAD method able to classify speech and the non-speech signal under noisy real-life conditions. The advantage of our proposed method is the ability to classify speech/non-speech under various kinds of noises and effectively determine speech signal under low-SNR.

VII. Conclusions

In this paper, we proposed an alternative approach of VAD method based on LSTM using modulation spectrum (MS) in addition to Mel-frequency cepstrum coefficients and pitch. The MS was calculated from the first coefficient of MFCC. We used the area under the curve of MS instead of the Q-value to represent the feature of MS. The MFCC, energy, pitch, and the feature of MS were aligned in one feature vector. The LSTM-based VAD was trained by using the noisy speech signals in three varieties of SNR. The evaluation was the comparison between the performance of the traditional method and that of our proposed method. We also evaluated the performance of our proposed method under unseen noises in real environments. The results showed that our proposed yield the improved accuracy up to 3.5% under real-life noisy condition. In the future work, other feature of MS such as the slope will be considered to improve the performance of the proposed method.

References

[1] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *IET Electronics Letters*, pp. 180-181,2000.

[2] J. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse condition: incidence on a DTW and HMM recognizer," *Secound European Conference on Speech Commmunication and Technology*,1991.

[3] J. Ramriez, J. Gorriz, and J. Segura, "Voice Activity Detection. Fundamentals and Speech Recognition System Robustness," *Robust Speech Recognition and Understanding*,pp. 1-22,2007.

[4] H. Hermansky, "Modulation Spectrum in Speech Processing," *Signal Analysis and Prediction*,pp. 395-406,1998.

[5] J. Sohn, N. Kim, and W. Sung, "A Statistical Model based Voice Activity Detection," *Signal Processing Letters*,vol6,no.1,pp. 1-3,1999.

[6] S. Tong, H. Gu, and K. Yu "A comparative study of robustness of deep learning approaches for vad," *International Conference on Acoustics, Speech, and Signal Processing,ICASSP*,2016.

[7] Y. Kanai, and M. Unoki, "Robust voice activity detection using empirical mode decomposition and modulation spectrum analysis," *International Symposium on Chinese Spoken Language Processing,ISCSLP*,2012.

[8] F. Eyben, F. Weninger, S. Squartini and B. Schuller, "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies," *International Conference on Acoustics, Speech, and Signal Processing,ICASSP*,2013.

[9] L. Atlas, S. Greenberg, and H. Hermansky, "The modulation spectrum and its application to speech science and technology," *Interspeech Tutorial*,Antwerp,Belgium,2007.

[10] D. Povey et al, "The Kaldi Speech Recognition toolkit," *IEEE Signal Processing Society*,2011.

[11] K. Kasuriya, V. Sornlertlamvanich, P. Chootrakool, N. Thatphithakkul, and C. Wutiwiwatchai, "Thai speech corpus for Thai speech recognition," *Oriental COCOSDA*,pp. 54-61,2003.

[12] H.G. Hirsch, "Aurora-5 Experimental Framework for the Performance Evaluation of Speech Recognition in Case of a Hands-free Speech Input in Noisy Environments," *http://aurora.hsnr.de/aurora-5*,2007.

[13] C. Fran et al, "Keras," *https://github.com/fchollet/keras/*.

[14] D.P. Kingma, and J. Ba "Adam: A Method for Stochastic Optimization," *the 3rd International Conference for Learning Representations*,2015.