*International Journal of Internet, Broadcasting and Communication Vol.11 No.4 26-30 (2019)*
*http://dx.doi.org/10.7236/IJIBC.2019.11.4.26*

**IJIBC 19-4-4**

# Voice Activity Detection Based on SNR and Non-Intrusive Speech Intelligibility Estimation

Soo Jeong An*, Seung Ho Choi**

*, **Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Seoul, Korea
* g_two@naver.com, **shchoi@snut.ac.kr

***Abstract***

*This paper proposes a new voice activity detection (VAD) method which is based on SNR and non-intrusive speech intelligibility estimation. In the conventional SNR-based VAD methods, voice activity probability is obtained by estimating frame-wise SNR at each spectral component. However these methods lack performance in various noisy environments. We devise a hybrid VAD method that uses non-intrusive speech intelligibility estimation as well as SNR estimation, where the speech intelligibility score is estimated based on deep neural network. In order to train model parameters of deep neural network, we use MFCC vector and the intrusive speech intelligibility score, STOI (Short-Time Objective Intelligent Measure), as input and output, respectively. We developed speech presence measure to classify each noisy frame as voice or non-voice by calculating the weighted average of the estimated STOI value and the conventional SNR-based VAD value at each frame. Experimental results show that the proposed method has better performance than the conventional VAD method in various noisy environments, especially when the SNR is very low.*

***Keywords:*** *Voice Activity Detection (VAD), SNR-based VAD, Non-intrusive speech intelligibility estimation, STOI, Deep neural network.*

## 1. INTRODUCTION

It is necessary to improve the performance of voice segment detection in the field of speech enhancement and communication. Conventional VAD methods estimate the voice segment by calculating SNR at each frame [1, 2]. There is a standard intrusive speech intelligibility estimation method, STOI (Short-Time Objective Intelligent Measure) [3] that is the method of calculating the correlation between the reference signal and distorted signal in the frequency domain. We cannot use STOI for VAD directly since there is no reference clean signal. Recently, the non-intrusive speech intelligibility estimation method based on deep neural network, which incorporating STOI values, was studied [4]. We incorporate non-intrusive speech intelligibility estimation method in order to improve the VAD performance. Therefore, we propose a new VAD method that uses a non-intrusive speech intelligibility estimation method as well as SNR-based

Manuscript Received: August. 18, 2019 / Revised: August. 21, 2019 / Accepted: August. 26, 2019
Corresponding Author: shchoi@seoultech.ac.kr
Tel: +82-02-970-6461, Fax: +82-02-979-7903
Dept. of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Korea

method.

## 2. VAD BASED ON SNR ESTIMATION

Conventionally, voice activity probability is obtained by estimating SNR at each frame as shown in Figure 1 [1, 2, 5]. Let x(t), d(t) and y(t) denote the speech, noise and noisy signal, respectively.

$$y(t) = x(t) + d(t) \tag{1}$$

The estimated SNR value for speech segment detection is calculated from the Equations (2) and (3) [2, 5].

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \quad : \quad \text{a priori signal to noise ratio} \tag{2}$$

$$\gamma_k \triangleq \frac{R_k{}^2}{\lambda_d(k)} \quad : \quad \text{a posteriori signal to noise ratio} \tag{3}$$

The noise variance $\lambda_d(k)$ and speech variance $\lambda_x(k)$ of the *k*th spectral component are estimated by using minimum statistics noise PSD estimation [6]. $R_k$ is the *k*th spectral component of the noisy signal. The likelihood ratio for the *k*th spectral component and the geometric mean of the likelihood ratios are given by Equations (4) and (5) [1, 5].

$$\Lambda_k = \frac{1}{1+\xi_k} \exp\left(\frac{\gamma_k \xi_k}{1+\xi_k}\right) \tag{4}$$

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \tag{5}$$

The value of the voice activity probability, V(n), at the *n*th frame can be obtained by the Equation (6) [5].

$$V(n) = \frac{1}{1+ e^{-\log \Lambda}} \tag{6}$$

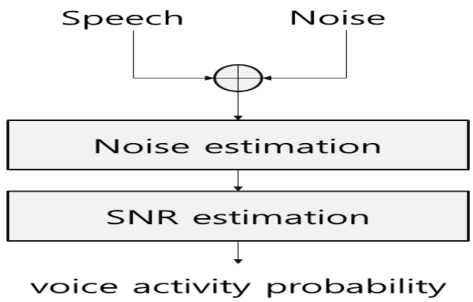


**Figure 1. SNR-based VAD method**

## 3. NON-INTRUSIVE SPEECH INTELLIGIBILITY ESTIMATION

As shown in Figure 2, the number of input nodes in deep neural network used for speech intelligibility estimation is equal to the order of feature vector [4]. For the training of the neural network, the value of the output is a frame-wise STOI score. In the test, we get the estimated speech intelligibility score per frame. We

use mel-frequency cepstral coefficient (MFCC) vectors in various noisy environments as shown in the figure [4,7]. In the test, the MFCC vector per frame enters the input of the neural network to obtain the estimated intelligibility value. The activation function of the neural network is the ReLU (Rectified Linear Units) [8] with an output of $\max(0, x)$ at input $x$. In addition, the ADAptive Moment estimation (ADAM), a statistical optimization algorithm, was used to learn neural network parameters [9].

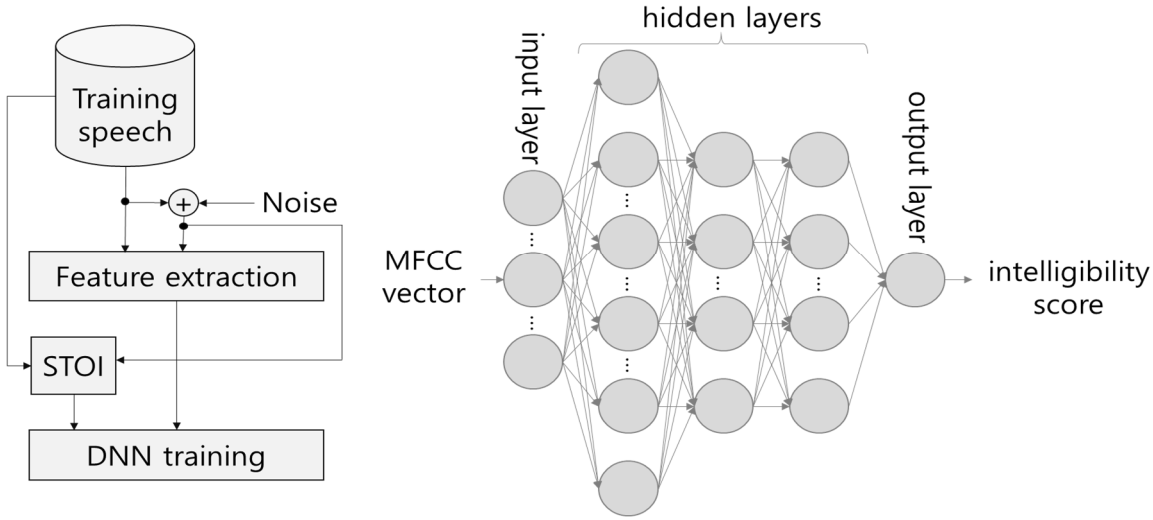


**Figure 1.Training and test structure for speech intelligibility estimation**

## 4. PROPOSED HYBRID VAD METHOD

We propose a new VAD method that uses the estimated speech intelligibility score I(n) at *n*th frame as well as the voice activity probability obtained from SNR estimation. We define a speech presence measure D(n) as in Equation (7) to classify each noisy frame as voice or non-voice, where the weights of V(n) and I(n) are controlled by the value of λ.

$$D(n) = \lambda V(n) + (1-\lambda) I(n) \tag{7}$$

## 5. EXPERIMENTS AND RESULTS

There are three hidden layers that have 1000, 400, 400 nodes each in the deep neural network for speech intelligibility estimation. NTT Korean Speech Database [10] was used for training and test. We used a 39-dimensional feature vector including 12 MFCCs and log energy, along with their delta and double delta values for the neural networks [4]. Neural networks were trained in 10 different environments: clean, noise, and noisy. The tests were conducted in 47 environments, including 37 environments not used in the training process. We used indoor noises as: Door, Bell ringing, Dog, Clock, Step noise. The SNR ranges from 0 to 20 dB for the noisy conditions. The Figure 3 is an example of the VAD result of reference, conventional and the proposed methods.

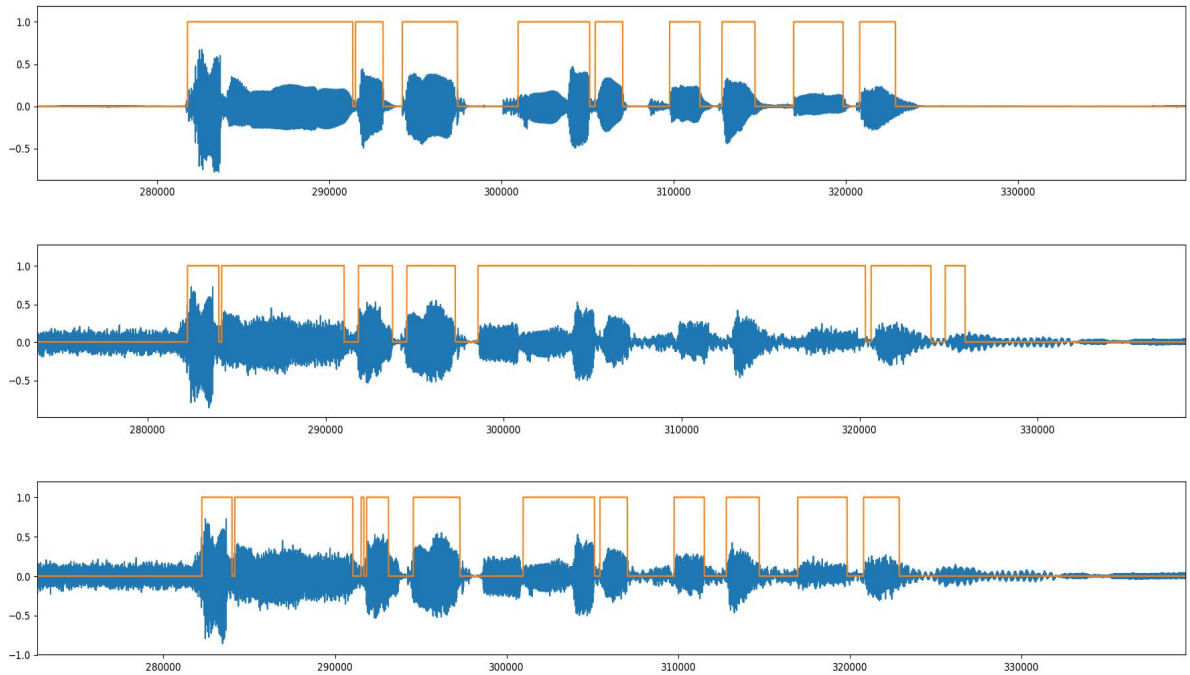


**Figure 2. An example of VAD results of reference, conventional and proposed methods**

Table 1 shows the comparison results of detection accuracy for each method. In the case of the hybrid method, the best accuracy was experimentally obtained when the λ is 0.2. The experimental results show that the proposed hybrid method is superior to the conventional method.

**Table 1. Comparison results of detection accuracy for each method**

| SNR [dB] | Accuracy [%] | |
|---|---|---|
| | SNR-based VAD | hybrid VAD |
| 0 | 69.84 | 98.19 |
| 10 | 77.56 | 98.35 |
| 20 | 82.25 | 98.89 |

## 4. CONCLUSION

We proposed the hybrid VAD method based on SNR and non-intrusive speech intelligibility estimation. The conventional VAD estimates the SNR to obtain a voice activity probability at each frame. For the estimation of speech intelligibility, the input and output of deep neural network are MFCC vector and frame-wise STOI score. We calculate the weighted average of the estimated STOI value and the conventional SNR-based VAD value at each frame. From the experiments in various noisy environments, we confirmed that the proposed hybrid method gives the superior detection performance to the conventional SNR-based VAD method.

## ACKNOWLEDGEMENT

This study was supported by the Research Program funded by the SeoulTech (Seoul National University of Science and Technology).

## REFERENCES

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, issue 1, pp. 1-3, Jan. 1999.
DOI: https://www.doi.org/10.1109/97.736233

[2] M. Vondrasek and P. Pollak, "Methods for Speech SNR estimation: Evaluation Tool and Analysis of VAD Dependency," *Radioengineering* 14(1), April 2005
DOI: https://doaj.org/article/a53fe518a9634318b417fb15a8c37fa8

[3] C.H. Taal, R.C. Hendrilks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.7, pp.2125–2136, 2011.
DOI: https://www.doi.org/10.1109/TASL.2011.2114881

[4] D. K. Yun, H. N. Lee, and S. H. Choi, "A Deep Learning-Based Approach to Non-Intrusive Speech Intelligibility Estimation," *IEICE Trans. Information and Systems*, pp. 1207-1208, Apr. 2018.
DOI: https://www.doi.org/10.1587/transinf.2017EDL8225

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 32, no. 6, Dec. 1984
DOI: https://www.doi.org/10.1109/TASSP.1984.1164453

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, July 2001
DOI: https://www.doi.org/10.1109/89.928915

[7] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum", *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 73-76, May 2001
DOI: https://www.doi.org/10.1109/ICASSP.2001.940770

[8] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines", *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.
DOI: https://dl.acm.org/citation.cfm?id=3104425

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv: 1412.6980, 2014.
DOI: https://arxiv.org/abs/1412.6980

[10] Multi-lingual speech database for telephonometry (1994). [Online]. Available: http://www.ntt-at.com/product/speech/. NTT Adv. Technol. Corp. Accessed 18 April 2016.