Dual Attention in Time and Frequency Domain for Voice Activity Detection

Joohyung Lee, Youngmoon Jung, Hoirin Kim

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, South Korea
{wngud701,dudans,hoirkim}@kaist.ac.kr

Abstract

Voice activity detection (VAD) is a challenging task in low signal-to-noise ratio (SNR) environment, especially in non-stationary noise. To deal with this issue, we propose a novel attention module that can be integrated in Long Short-Term Memory (LSTM). Our proposed attention module refines each LSTM layer’s hidden states so as to make it possible to adaptively focus on both time and frequency domain. Experiments are conducted on various noisy conditions of Aurora 4. Our proposed method obtains the 95.58% accuracy under the ROC curve (AUC), achieving 22% relative improvement compared to baseline, with only 2.44% increase in the number of parameters. Besides, we utilize a focal loss for alleviating the performance degradation caused by imbalance between speech and non-speech sections in training sets. The results show that the focal loss can improve the performance in various imbalance situations compared to the cross entropy loss, a commonly used loss function in VAD.

Index Terms: voice activity detection, long short-term memory, attention, class imbalance, focal loss

1. Introduction

Voice activity detection (VAD) is a kind of binary classification which classifies a frame into speech or non-speech. It is an important pre-processing step in speech application like automatic speech recognition (ASR), speaker recognition, speech enhancement and speech coding, etc.

The early approaches to VAD were based on signal processing-based approaches using time-domain power [1], zero crossing rate (ZCR) [2], cepstral features [3] or spectral entropy [4]. Besides, statistical model-based approaches have been widely adopted using Gaussian models for speech and noise signals [5][6].

Recently as deep learning has been on the rise, it has shown its effectiveness on finding the optimal VAD models such as deep neural networks (DNNs) [7][9], convolutional neural networks (CNNs) [10][11], Long Short-Term Memories (LSTMs) [11][15], and the combination of deep neural networks [16][17].

Despite those deep learning-based VAD models have achieved marked improvements, VAD is still a challenging task in low signal-to-noise ratio (SNR) environments. To improve the robustness in noisy environment, we propose a novel VAD model based on attention method. Our architecture is motivated by the attention module integrated to CNN architecture used in computer vision field [18][19]. Especially, our proposed method is mainly based on convolutional block attention module (CBAM) [19].

Meanwhile, in supervised learning-based classification problem, class imbalance of training data can bring about deterioration since easily classified samples dominate the training procedure [20][21]. In case of VAD as well, audio samples in database usually show the imbalance between speech and non-speech sections. Indeed, cross entropy loss, broadly used in VAD, is not suitable for handling the class imbalance. On the other hand, a focal loss proposed in [22] has a modulating term which is able to focus learning on minor samples in class imbalance situations. In the experiment, we utilize the focal loss in various class imbalance situations and prove that it is conducive to class imbalance situation for VAD.

The remaining part of paper is organized as follows. Section 2 describes 4 types of proposed attention modules. Section 3 indicates problems about class imbalance in VAD and compares focal loss with cross entropy loss. Section 4 describes the experimental setup and Section 5 shows the results and analysis of experiments. Then, we conclude this work in Section 6.

2. Attention Module

CBAM refines the 3-dimensional feature map by combining channel attention module and spatial attention module in CNN architecture [19]. Motivated by CBAM, several studies have been conducted in speech processing field [23][24]. However, to our best knowledge, there was no attempt to apply CBAM for VAD. Unlike original CBAM, we change the backbone architecture from CNN to LSTM. The reason of using LSTM instead of CNN is that LSTM shows the best performance among DNN, CNN and LSTM in VAD with a similar number of parameters [25][26].

The structure of proposed attention-based LSTM model is shown in Figure 1. $X \in \mathbb{R}^{T \times I}$ is input acoustic features for model. $T$ denotes the length of time steps (sequence length) and $I$ denotes the dimension of acoustic features. When $X$ is fed into the first LSTM layer (LSTM 1), then hidden states $H \in \mathbb{R}^{T \times D}$ with $D$ hidden nodes are drawn. In basic LSTM, hidden states $H$ are fed to the next LSTM layer (LSTM 2) directly [25]. But in this paper, hidden states are refined by proposed attention module before being fed to the next LSTM layer (LSTM 2). For refining hidden states, we regard hidden states $H$ as a kind of 2-dimensional hidden feature map. In following subsections, we propose 4 kinds of attention modules.

Figure 1: Illustration of proposed attention-based LSTM model.
2.1. Temporal Attention (TA)

The temporal attention (TA) exploits temporal information and concentrates on specific time steps for improving model’s ability to discriminate speech frames from non-speech frames. Figure 2a illustrates the process to obtain $H_{\text{temp}}$, output of TA module. The hidden feature map $H$ is pooled in three ways, max, average and standard deviation pooling along the frequency axis that can be represented as $H_{\text{temp}}^\text{max}, H_{\text{temp}}^\text{avg}, H_{\text{temp}}^\text{std} \in \mathbb{R}^{T \times D}$, respectively (step 1). These three pooled feature vectors are concatenated and being convolved by 1-dimensional convolution layers (step 2 - 4). $H_{\text{temp}}$, the output of last convolution layer, is expanded (copied) as original hidden feature map $H$ and activated by sigmoid function. Finally, it is merged to $H$ by element-wise summation, then refined hidden feature map $H'$ is obtained. TA can be represented as below.

\[
H_{\text{temp}} = f_{\text{temp}}^\text{11}(H_{\text{temp}}^\text{max}, H_{\text{temp}}^\text{avg}, H_{\text{temp}}^\text{std}) \quad (1)
\]

\[
H' = H \oplus \sigma(H_{\text{temp}}) \quad \text{for} \quad H_{\text{temp}} \in \mathbb{R}^{T \times D}, \quad (2)
\]

where $f_{\text{temp}}^\text{11}$ denotes the 1-dimensional convolution with 11 of kernel size in TA module. It consists of 4 layers and the number of filters is 3, 5, 5 and 1, respectively. $\sigma$ denotes the sigmoid function and $\oplus$ denotes the element-wise summation.

2.2. Frequential Attention (FA)

The frequential attention (FA) is same with TA but for pooling direction and kernel size of convolution layer. Figure 2b illustrates the process to obtain $H_{\text{freq}}$, output of FA module. $H_{\text{freq}}^\text{max}, H_{\text{freq}}^\text{avg}, H_{\text{freq}}^\text{std} \in \mathbb{R}^{T \times D}$ are the max, average and standard deviation pooling results of hidden feature map $H$ along the time axis (step 3). Like in TA, these three feature vectors are concatenated, passed to convolution layers (step 2 - 4) and sigmoid function. Also, after being expanded as $H$, it is merged to $H$ by element-wise summation for obtaining refined hidden feature map $H'$. FA can be represented as below.

\[
H_{\text{freq}} = f_{\text{freq}}^\text{11}(H_{\text{freq}}^\text{max}, H_{\text{freq}}^\text{avg}, H_{\text{freq}}^\text{std}) \quad (3)
\]

\[
H' = H \oplus \sigma(H_{\text{freq}}) \quad \text{for} \quad H_{\text{freq}} \in \mathbb{R}^{T \times D}, \quad (4)
\]

where $f_{\text{freq}}^\text{11}$ denotes the 1-dimensional convolution with 21 of kernel size in FA module. Sequence length $T$ in training steps is fixed by predefined value to build the mini-batch of training. But that in testing steps is variable according to audio sample’s utterance. This mismatch of sequence length causes the disparate tendency of pooled values in both steps with degradation of performance. To circumvent this problem, in testing steps, utterances are divided by sequence length $T$, which is defined in training steps, then FA is applied to each divided segments. For example, if sequence length $T$ of training steps is 50 time steps, FA is applied every 50 time steps in test data, e.g. 1-50 time steps, 51-100 time steps, etc. The last left over steps are pooled by themselves.

2.3. Dual Attention 1 (DA-1)

To exploit both temporal and frequential information simultaneously, the dual attention 1 (DA-1) is suggested. The process of DA-1 is illustrated in Figure 2a. DA-1 uses hidden feature map $H$ directly and convolving it by 2-dimensional convolution layers (step 1 - 2). Merging method is element-wise summation like in TA and FA.

\[
H_{\text{dual}} = f_{\text{dual}}^\text{7\times7}(H) \quad (5)
\]

\[
H' = H \oplus \sigma(H_{\text{dual}}) \quad \text{for} \quad H_{\text{dual}} \in \mathbb{R}^{T \times D}, \quad (6)
\]

where $f_{\text{dual}}^\text{7\times7}$ denotes the 2-dimensional convolution in the DA-1 module with kernel size of 7. It consists of 3 layers and the number of filters is 1, 3 and 1, respectively.

2.4. Dual Attention 2 (DA-2)

The other way for exploiting both temporal and frequential information is using $H_{\text{temp}}$ and $H_{\text{freq}}$ at once in summation, called dual attention 2 (DA-2), which is the combination of TA and FA. The activation function and merging method are same as in TA and FA.

\[
H' = H \oplus \sigma(H_{\text{temp}} \oplus H_{\text{freq}}). \quad (7)
\]

Computations for obtaining $H_{\text{temp}}$ and $H_{\text{freq}}$, in Eq. (1) and Eq. (3), are executed in parallel.

Every convolution operation in proposed attention modules is followed by batch normalization and ReLU activation function. However, in the very last layer of attention module, batch normalization and activation function is not used because of using sigmoid function before merging. Attention modules are applied after every hidden feature map, even for hidden feature map from last LSTM layer. Also, same attention module is shared across all hidden feature maps from different LSTM layers. It means there is no need to train several attention modules upon the number of hidden layers in LSTM.

3. Loss Functions

Since it is hard to record audio samples in equal or similar ratio of speech to non-speech, imbalance between speech and non-speech sections can be found easily in lots of datasets. To balance the ratio of speech to non-speech, many researchers manipulate the data by artificially appending silence segment at the beginning and the end of audio samples in training datasets [9, 14, 28-32]. To avoid this inconvenience, we utilize the focal loss, revised version of cross entropy loss [22].
3.1. Cross Entropy Loss

Cross entropy loss is the entropic-measure of discrepancy across distributions, represented as below.

\[
I_{CE}(y_t) = - \log(y_t) \tag{8}
\]

\[
y_t = \begin{cases} 
\hat{y} & \text{if } y_t = 1 \\
1 - \hat{y} & \text{otherwise}
\end{cases}
\tag{9}
\]

where \( y \) is label and \( \hat{y} \) is model’s predicted probability for label \( y = 1 \). Thanks to its convexity in optimization, it is widely used in deep learning-based task. In spite of its usefulness, cross entropy loss is hard to naturally handle the inefficient training caused by class imbalance.

3.2. Focal Loss

To mitigate the inefficient training in class-imbalanced environment, focal loss is suggested and described as below.

\[
l_{FL}(y_t) = -(1 - y_t)^\gamma \log(y_t), \tag{10}
\]

where \( \gamma \) is tunable parameter named focusing parameter and \( y_t \) is same with in cross entropy loss, Eq. (9). The main difference between cross entropy loss and focal loss is modulating factor, \((1 - y_t)^\gamma\). Modulating factor is increased when the gap between target and predicted value is increased (misclassified case). Otherwise, when the gap is decreased, modulating factor is also decreased (well-classified case). From these mechanism, it strengthens the significance of correcting misclassified examples and alleviates the bias oriented to dominating class.

4. Experimental Setup

4.1. Datasets

The experiments were conducted on Aurora 4 [33] which contains 7138 and 330 clean utterances for training and testing, respectively. All the clean utterances of training data were corrupted by the public 100 noise types\(^1\) at SNR from -10 to 15 dB in 5 dB steps. Noise types and SNRs were selected randomly. This procedure was repeated until training sets reached about 60 hours long. To evaluate the performance in mismatched noisy conditions, we added 5 unseen noises (babble, destroyer-engine, F16 cockpit, factory and street) with 4 SNRs (-5, 0, 5 and 10 dB) to all of testing data. Because Aurora 4 data show speech dominated class imbalance, 1 second of silence were inserted at back and forth of each utterance in training sets (1 sec padding).

To do experimental work for focal loss, we manipulated training sets for making various imbalance situations. At first, a kind of endpoint detection was executed based on ground-truth (EPD). That is to say, the front part before first speech frame and the latter part after last speech frame were deleted. Secondly, no manipulation was conducted (no padding). For making opposite condition, we inserted the silence at back and forth of audio samples for 2 seconds and 3 seconds (2 sec padding and 3 sec padding, respectively). The focusing parameter \( \gamma \) of focal loss in Eq. (10) was set as 0.2, 0.4, 0.6, 0.8, 1.0, 2.0 and 3.0.

4.2. Setting

40-dimensional log Mel-filterbanks were used as acoustic features with 25ms frame length and 10ms shift length. The ground-truth of noisy speech was extracted by applying Sohn\footnote{web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html}.

5. Results

5.1. Comparison of different attention modules

Table 2 represents the results of the baseline (LSTM\(_{64}\)) and baselines integrated with all of proposed attention modules. Evaluation metric is the area under the ROC curve (AUC) [14]. The results of 5 noises are averaged along same SNR level and the number of parameters is also compared.

From this table, we can observe that all of attention-based models outperform the baseline. And it would be quite proper to say that attention even only for single domain, time or frequency, can help LSTM model to be optimized in VAD. In -5 dB SNR, the frequentational attention (FA) in Section 2.2 outperforms the temporal attention (TA) in Section 2.1. However, both show similar results in other SNR levels. It implies that attention in frequency domain is more effective than in time domain especially in desperately noisy environment under 0 dB SNR.

The dual attention 2 (DA-2) in Section 2.4 shows the best results throughout whole SNR levels. It is natural result because DA-2 utilizes both of temporal and frequentational information. However, although the dual attention 1 (DA-1) in Section 2.3 also utilizes both information, it shows the lowest result among the attention modules. It means that DA-2 uses both information more precisely than DA-1 by using 1-dimensional convolution separately in each domain.

However, DA-1 outperforms the baseline as well and uses

Table 2: Averaged AUC(%) of 5 noises and number of parameters. In all tables in this paper, the best results are highlighted in bold font and RI with parenthesis represents the relative improvement (except for Table 4).

| SNR  | LSTM\(_{64}\) | w/ TA | w/ FA | w/ DA-1 | w/ DA-2 |
|------|---------------|-------|-------|---------|---------|
| -5 dB| 87.05         | 88.37 | 89.38 | 88.33   | 90.06   |
| 0 dB | 94.13         | 94.92 | 94.89 | 94.85   | 95.42   |
| 5 dB | 97.42         | 97.77 | 97.67 | 97.77   | 97.90   |
| 10 dB| 98.74         | 98.82 | 98.83 | 98.88   | 98.93   |
| Avg  | 94.33         | 94.97 | 95.19 | 94.96   | 95.58   |

RI (1.11 %) (11.17 %) (9.67 %) (22.05 %) (22.05 %)

| # Parim | 95.809        | 96.627 | 97.24 | 96.565 | 98.14 | 95.88 |
| (Increase) | (-) (0.85%) (1.58 %) (0.79 %) (2.44 %) (9.80 %)|
Table 4: Averaged AUC(%) of 5 noises in all SNRs for the baseline (LSTM_64) and DA-2 based model. CE and FL denote cross entropy and focal loss, respectively. Value in parenthesis after FL is the focusing parameter $\gamma$. Results which outperform the CE-base result are highlighted in bold font. The bottom row represents the ratio of speech (S) to non-speech (NS) of training data in each situation.

| Loss  | 1 sec padding | 2 sec padding | 3 sec padding |
|-------|---------------|---------------|---------------|
|       | w/ DA-2 | Baseline | w/ DA-2 | Baseline | w/ DA-2 | Baseline | w/ DA-2 | Baseline | w/ DA-2 | Baseline | w/ DA-2 |
| EPD   |            |            |            |            |            |            |            |            |            |            |            |
| CE    | 91.70      | 93.45      | 92.64      | 94.81      | 94.33      | 95.58      | 94.47      | 95.42      | 94.38      | 95.33      |            |
| FL (0.2) | 92.33    | 93.82      | 93.30      | 94.64      | 94.40      | 95.43      | 94.64      | 95.49      | 94.53      | 95.23      |            |
| FL (0.4) | 92.33    | 93.33      | 93.16      | 94.04      | 94.39      | 95.39      | 94.66      | 95.50      | 94.61      | 95.22      |            |
| FL (0.6) | 92.27    | 93.34      | 93.04      | 94.04      | 94.40      | 95.52      | 94.66      | 95.41      | 94.41      | 95.49      |            |
| FL (0.8) | 92.23    | 93.44      | 92.91      | 95.40      | 94.39      | 95.59      | 94.55      | 95.41      | 94.47      | 95.46      |            |
| FL (1)  | 92.16     | 92.88      | 92.55      | 94.88      | 94.39      | 95.57      | 94.53      | 95.48      | 94.40      | 95.35      |            |
| FL (2)  | 91.78     | 92.62      | 92.31      | 94.54      | 94.29      | 95.50      | 94.27      | 95.21      | 94.18      | 95.19      |            |
| FL (3)  | 91.38     | 92.19      | 92.19      | 93.95      | 94.09      | 95.52      | 94.20      | 94.56      | 93.99      | 94.97      |            |

Ratio (S / NS) | 69.96 / 38.76 | 62.14 / 38.76 | 48.59 / 51.41 | 40.23 / 59.77 | 34.32 / 65.68

Table 3: Averaged AUC(%) of 5 noises and the number of parameters for other baseline models and dual attention 2 (DA-2).

| Model | Avg. (RI) | # Param. (Increase) |
|-------|-----------|---------------------|
| Baseline | LSTM_64  | 94.42 (+)          | 205.121 (+) |
|         | CLDNN_64 | 94.53 (+)          | 129.883 (+) |
|         | CLDNN_20 | 94.55 (+)          | 215.927 (+) |
| Attention | LSTM_64  | 95.36 (17.20 %)   | 207.457 (1.14 %) |
|         | CLDNN_64 | 95.55 (18.65 %)   | 132.219 (1.30 %) |
|         | CLDNN_20 | 95.42 (15.96 %)   | 218.263 (1.08 %) |

Figure 3: The last hidden layer’s feature map of 20 frames from 446c0201.wav: baseline and dual attention-based model.

The least number of parameters among the attention modules. Also, the increase in number of parameters in all of attention modules is under 2.5 %, which is negligible.

For showing effectiveness of attention, test waveform sample 446c0201.wav, corrupted by F16 cockpit with 0 dB SNR, was selected. Figure 3 shows the hidden feature map of last LSTM layer from baseline in left column and the hidden feature map of last LSTM layer from DA-2 based model in right column. The upper graphs represent ground-truth and the lower graphs represent predicted results from randomly selected consecutive 20 frames in the sample (1:speech / 0:non-speech). DA-2 based model concentrates on time steps of speech frame and suppress time steps of non-speech frame by TA (indicated by the red rectangular). In addition, differences of each hidden units are more distinct in DA-2 based model (indicated by the blue rectangular). It means DA-2 can strengthen the specific hidden units in helpful way to improve the model’s ability. As a result, DA-2 based model shows more accurate prediction rather than baseline.

The results of expanding experiments to other 3 baseline models are reported in Table 3. We find that DA-2 module can improve the performance in all of baselines, even in CLDNN. Also, the number of parameters is increased in little in all 3 cases. It means DA-2 module can be flexibly integrated to any LSTM model with a small increase in number of parameters.

Table 5: Mean and standard deviation of the top row in Table 4

| Model | Mean (RI) | Standard deviation (Increase) |
|-------|-----------|-------------------------------|
| Baseline | LSTM_64  | 93.50 (-)                     |
| Attention (w/ DA-2) | 94.92 (21.85 %) | 1.13 (+) |

5.2. Focal loss for various imbalance situations

Table 4 describes the results of experiment about focal loss. LSTM_64 was used as baseline and compared with DA-2 based model. The bottom row of table represents the ratio of speech to non-speech in each situation. Ratio of speech is decreased as column is moved from left (EPD) to right (3 sec padding). It can be found that focal loss is effective in all of situations even for balanced situation (1 sec padding). Interestingly, in speech dominated situation (EPD and no padding), focal loss shows more improved results than in opposite situations (2 sec padding and 3 sec padding). Also, the effect of focal loss is less remarkable in DA-2 based model than baseline generally.

Table 5 represents the mean and standard deviation of cross entropy-based results (the top row in Table 4) from 5 different padding situations. For comparing baseline and DA-2 based model mathematically in imbalance situations, statistical values are obtained along same model. As we can see, mean of DA-2 based model is higher than baseline with 21.85 % of relative improvement. Besides, in terms of standard deviation, fluctuation of performance upon class imbalance is less shown up in DA-2 based model. It can be said that DA-2 module can mitigate the drop of performance in imbalance situation. Also, focal loss can be the countermeasures for imbalance situation in VAD task, especially in speech dominated situation.

6. Conclusion

This paper proposed a novel VAD applying dual attention module which exploits the time and frequency information and infers optimal attention vectors for each domain. As a result, the proposed attention module improves the performance compared to baseline in unseen noise environment with a slight increase in number of parameters. In addition, the proposed attention module can be flexibly integrated to other LSTM-based baselines for better performance. By using focal loss in diverse imbalance situations, performance degradation is alleviated compared to using cross entropy loss.

7. Acknowledgements

This material is based upon work supported by the Ministry of Trade, Industry and Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No.10063424, Development of distant speech recognition and multi-task dialog processing technologies for in-door conversational robots).
8. References

[1] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” The Bell System Technical Journal, vol. 54, no. 2, pp. 291–315, 1975.

[2] J. C. Junqua, B. Reaves, and B. Mark, “A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognition,” EUROspeech, pp. 1371–1374, 1991.

[3] J. H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” IEEE Transactions on Signal Processing, vol. 54, no. 6, pp. 1965–1976, 2006.

[4] J. Shen, J. Hung, and L. Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments,” in Proc. of International Conference on Spoken Language Processing (ICSLP), 1998, pp. 232–235.

[5] 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, 1999.

[6] J. H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” IEEE Transactions on Signal Processing, vol. 54, no. 6, pp. 1965–1976, 2006.

[7] J. Shen, J. Hung, and L. Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5695–5699.

[8] 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, 1999.

[9] J. H. Chang, N. S. Kim, and S. K. Mitra, “Voice activity detection based on multiple statistical models,” IEEE Transactions on Signal Processing, vol. 54, no. 6, pp. 1965–1976, 2006.

[10] J. Shen, J. Hung, and L. Lee, “Robust entropy-based endpoint detection for speech recognition in noisy environments,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1999, pp. 1263–1284, 2009.

[11] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proc of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.

[12] S. Yang and A. Dai, “Frequency and temporal convolutional attention for text-independent speaker recognition,” arXiv e-prints, p. arXiv:1910.07364, 2019.

[13] S. Shi, Q. Huang, and T. Hain, “Robust speaker recognition using speech enhancement and attention model,” arXiv e-prints, p. arXiv:2001.05031, 2020.

[14] S. Tong, H. Gu, and Y. Yu, “A comparative study of robustness of deep learning approaches for VAD,” in Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5695–5699.

[15] M. Wang, Q. Huang, J. Zhang, Z. Li, H. Pu, J. Lei, and L. Wang, “Deep learning approaches for voice activity detection,” in Proc. of the International Conference on Cyber Security Intelligence and Analytics (CSIA), 2019, pp. 816–826.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] S. Graf, T. Herbig, M. Buck, and G. Schmidt, “Features for voice activity detection: a comparative analysis,” EURASIP Journal on Advances in Signal Processing, 2015, no. 91, pp. 1–15, 2015.

[18] T. Drugman, Y. Stiliou and, and Y. K. M. Akamine, “Voice activity detection: merging source and filter-based information,” IEEE Signal Processing Letters, vol. 23, no. 2, pp. 252–256, 2016.

[19] I. C. Yoo, H. Lim, and D. Yook, “Formant-based robust voice activity detection,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2238–2245, 2015.

[20] P. K. Ghosh, A. Tsiartas, and S. Narayanan, “Robust voice activity detection using long-term signal variability,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 600–613, 2011.

[21] G. Ameenja and B. Yeegnanarayana, “Single frequency filtering approach for discriminating speech and nonspeech,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 705–717, 2015.

[22] N. Parihar and J. Picone, “Aurora working group: DSR front end LVCSR evaluation AU38402,” Inst. for Signal and Information Process, Mississippi State University, Tech. Rep, vol. 40, p. 94, 2002.

[23] A. Hanley and J. McNeil, “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve,” Radiology, vol. 143, pp. 29–36, 1982.