Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 12, No. 3, 2024, 337-347.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Deep Learning Approach for Robust Voice Activity Detection: Integrating CNN and Self-Attention with Multi-Resolution MFCC

Khadijeh Aghajani*

Department of computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran.

Article Info

Article History:
Received 27 July 2024
Revised 02 October 2024
Accepted 02 November 2024

DOI:10.22044/jadm.2024.14839.2582

Keywords:
Voice Activity Detection, Self-Attention Mechanism, Multi-Resolution Mel-Frequency Cepstral Coefficients, Deep Learning.

**Corresponding author: kh.aghajani@umz.ac.ir (Kh. Aghajani).*

Abstract

Voice Activity Detection (VAD) plays a vital role in various audio processing applications, such as speech recognition, speech enhancement, telecommunications, satellite phone, and noise reduction. The performance of these systems can be enhanced by utilizing an accurate VAD method. In this paper, multiresolution Mel-Frequency Cepstral Coefficients (MRMFCCs), their first and second-order derivatives (delta and delta2), are extracted from speech signal and fed into a deep model. The proposed model begins with convolutional layers, which are effective in capturing local features and patterns in the data. The captured features are fed into two consecutive multi-head self-attention layers. With the help of these two layers, the model can selectively focus on the most relevant features across the entire input sequence, thus reducing the influence of irrelevant noise. The combination of convolutional layers and self-attention enables the model to capture both local and global context within the speech signal. The model concludes with a dense layer for classification. To evaluate the proposed model, 15 different noise types from the NoiseX-92 corpus have been used to validate the proposed method in noisy condition. The experimental results show that the proposed framework achieves superior performance compared to traditional VAD techniques, even in noisy environments.

1. Introduction

Voice Activity Detection (VAD) is a vital phase in many speech processing-based applications such as automatic speech recognition, speech enhancement, telecommunications, speech encoding, satellite phone, speaker verification, and noise reduction systems. As an example, in a speaker verification task, non-speech intervals in speech files lack speaker information. So, VAD can be used to identify active speech segments before the feature extraction process [1].

The primary aim of VAD is to accurately distinguish between speech/non-speech segments in an audio signal, especially in noisy environments. It determines the starting/terminating points of active speech which significantly enhances the performance of subsequent processing stages.

Traditional VAD methods were based on heuristic methods or simple machine learning approaches using handcrafted features, such as short time energy level [2,3], zero-crossing rate [3], spectral information [4], Mel frequency cepstral coefficients (MFCC) [5], Cochleagrams [6], and wavelet energy [7,8]. These methods were very simple and computationally efficient but usually failed in challenging noisy environments. To improve the robustness of such methods, statistical model-based approaches, such as Gaussian Mixture Models (GMM) [9,10,11] and Hidden Markov Models (HMM) [12,13], were introduced to model speech and noise features using an appropriate probabilistic model. In [14], the statistical characteristics of the sub-band temporal envelope and the sub-band long-term signal variability are computed. By combining the two features the VAD decision according to the fusion decision has been made. In total, the traditional methods struggle in noisy environments, leading to reduced accuracy in detecting speech. These techniques typically rely on fixed thresholds or assumptions about signal distribution, making them less adaptable to varying noise conditions.

These days, the use of deep learning tools has brought significant progress in VADs, as in many other speech processing applications. In [15, 16, 6], taking into account multiple hand-crafted features, a deep multi-layered perceptron model has been used to detect the active area of the audio signals. Various combinations of Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and dense networks, have been used for VAD. They show promising results compared to traditional approaches due to their superior feature extraction and classification capabilities. Jung et al. have proposed a method called self-adaptive soft VAD which incorporates a deep neural network (DNN)-based VAD into a deep speaker embedding system [17].

In [18, 19, and 20], by considering Log-Mel energy spectrogram as an input to CNN layers followed by a dense layer, detection has been performed. Vecchiotti et al. have presented a CNN-based model for joint speech detection and speaker localization according to Log-Mel and GCC-PHAT information [21].

In [22], a combination of the CNN-based method and the MLP is used for VAD. In this method, the input signal is passed through Conv1D layers, then the extracted feature-map is concatenated to some handcrafted extracted features, and finally, after passing through a dense layer, the conclusion is made. In [23], the aim is improving VAD in noisy conditions. Their proposed method was a combination of CNN model and a De-noising Auto-encoder (DAE), by considering acoustic features and their delta features in noisy conditions. Wilkinson and Niesler proposed an end-to-end architecture consisting of both convolutional neural network and bidirectional long short-term memory (Bi-LSTM) for voice activity detection [24]. They utilized sequences of 32×32 spectrogram images as input to the model. Similar to [24], the combination of CNN and LSTM has been used in [25, 26, and 27] to detect active speech area. In [26], a method called CLDNN (Convolutional, Long Short-Term Memory, Deep Neural Networks) has been introduced. The raw waveform has been considered as the model’s input instead of Log-Mel features. They came to the conclusion that by using CNNs, better features can be extracted from raw data than other features such as Log-Mel. Jia et al. proposed a method called MagicNet incorporating the MobileNet and gated recurrent neural network (GRU) [28]. To reduce the number of parameters, the CNN layers are constructed with 1D depth-wise separable convolutions and a residual architecture. In [29], three features—Mel-frequency Cepstral Coefficients (MFCC), log filter banks, and spectral subband centroid—are extracted, fused, and fed into a classifier comprising three Recurrent Neural Networks (RNNs) with 256 neurons each and a fully connected layer with two neurons.

In [30], the VAD decision was derived from a simple Long Short-Term Memory (LSTM) network trained on auditory speech features, including energy, zero-crossing rate (ZCR), and 13th-order Mel Frequency Cepstral Coefficients (MFCC). In [31], a combination of Convolutional Neural Network followed by a Self-Attention (SA) Encoder has been proposed. Their proposed method was capable of processing the entire signal at once. In [32], with the aim of reducing the computational complexity of speaker diarization, the attention system of a speaker embedding extractor has been established as a weakly supervised VAD model.

In this paper, a novel VAD method that utilizes multiresolution Mel-Frequency Cepstral Coefficients has been proposed. This approach captures speech characteristics at multiple resolutions, ensuring a more detailed representation of the signal. To achieve robustness across noisy conditions, a deep model incorporating convolutional layers, followed by two multi-head attention layers is proposed. Convolutional layers can effectively extract spatial features from the input audio signals. Moreover, multi-head attention mechanism can enhance the model's ability to focus on relevant features. It can help the model to dynamically prioritize important features, especially in noisy environments, where distinguishing speech from noise is crucial. Finally, a dense layer performs the classification task. The proposed framework is evaluated considering 15 different noise types from the NoiseX-92 corpus, demonstrating its robustness and effectiveness in various noisy environments.

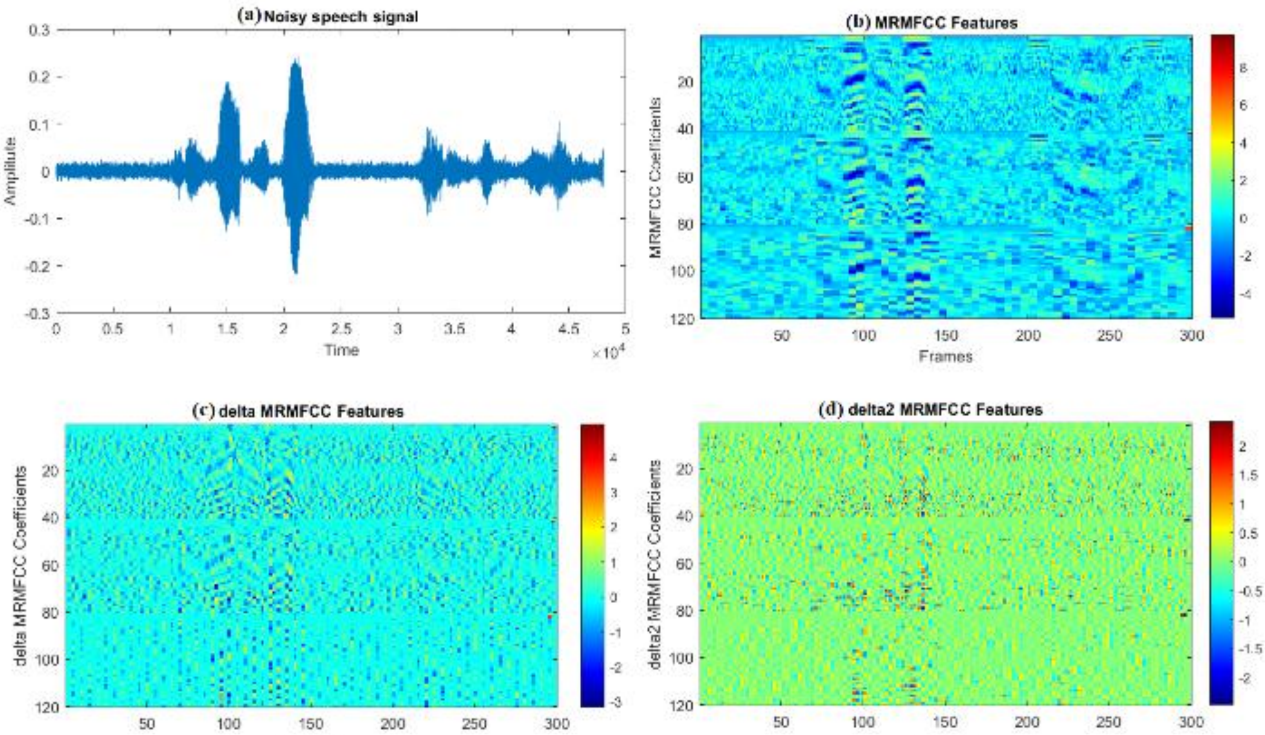


Figure 1. (a) Sample noisy speech signal. (b), (c), and (d) are MRMFCCs, delta MRMFCCs, and delta2 MRMFCCs, respectively. In each channel, bands 1 to 40, 41 to 80, and 81 to 120 are dedicated to high resolution, medium resolution, and low resolution coefficients, respectively.

With the proposed method, VAD is enhanced in terms of robustness and adaptability, especially in noisy environments. This makes it well-suited for various real-world applications. In telecommunications, it reduces the transmission of silence and conserves bandwidth by accurately detecting speech segments. In rescue operations, it can be crucial for detecting human voices amidst noise. For individuals with hearing impairments, the system can improve speech quality by filtering out noise. Lastly, it can enhance the performance of intelligent voice-command systems by distinguishing speech in noisy environments.

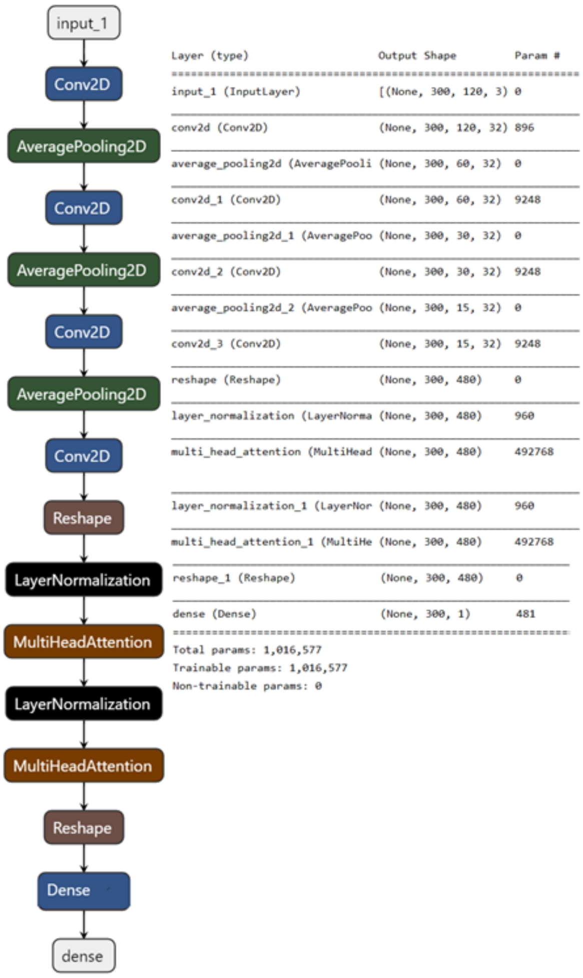
The rest of this paper is organized as follows: Section 2 describes the proposed framework, Section 3 presents the experimental results, and Section 4 concludes the paper with a conclusion.

2. Proposed method

In this research, a 3D audio image is generated by extracting multiresolution Mel-Frequency Cepstral Coefficients (MRMFCCs) along with their first and second-order derivatives from the speech signal. The use of MRMFCCs allows the model to capture detailed frequency variations at multiple resolutions, making it robust in noisy conditions. The obtained audio images are fed into a deep CNN-based model. The model architecture combines convolutional layers for efficient feature extraction with self-attention layers to capture long-range temporal dependencies, which are crucial for distinguishing speech from noise. The integration of CNNs with attention mechanisms enhances the model's ability to focus on the most relevant features, ensuring accurate detection across varying noise environments. This choice of architecture was motivated by the need for a model that can generalize well to diverse acoustic conditions, offering improved noise robustness and speech-non-speech discrimination. To evaluate the proposed method under real-world conditions, data augmentation using different noise types has been performed. The proposed framework is detailed below.

2.1. Data generation

To develop and validate the proposed framework, the TIMIT dataset has been employed [33]. This dataset is a widely used speech corpus designed for acoustic-phonetic analysis and speech recognition research. It consists of 630 speakers from 8 different dialect regions in the United States, with each speaker reading up to 10 phonetically rich sentences. The dataset provides a balanced and diverse range of speech samples in terms of speaker accent, gender, and speaking style, making it suitable for speech-related tasks such as Voice Activity Detection (VAD). All audio files were sampled at 16 kHz.



```
Layer (type)                 Output Shape         Param #
=================================================================
input_1 (InputLayer)         [(None, 300, 120, 3) 0
_________________________________________________________________
conv2d (Conv2D)              (None, 300, 120, 32) 896
_________________________________________________________________
average_pooling2d (AveragePooli (None, 300, 60, 32)  0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 300, 60, 32)  9248
_________________________________________________________________
average_pooling2d_1 (AveragePoo (None, 300, 30, 32)  0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 300, 30, 32)  9248
_________________________________________________________________
average_pooling2d_2 (AveragePoo (None, 300, 15, 32)  0
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 300, 15, 32)  9248
_________________________________________________________________
reshape (Reshape)            (None, 300, 480)     0
_________________________________________________________________
layer_normalization (LayerNorma (None, 300, 480)     960
_________________________________________________________________
multi_head_attention (MultiHead (None, 300, 480)     492768

_________________________________________________________________
layer_normalization_1 (LayerNor (None, 300, 480)     960
_________________________________________________________________
multi_head_attention_1 (MultiHe (None, 300, 480)     492768
_________________________________________________________________
reshape_1 (Reshape)          (None, 300, 480)     0
_________________________________________________________________
dense (Dense)                (None, 300, 1)       481
=================================================================
Total params: 1,016,577
Trainable params: 1,016,577
Non-trainable params: 0
```

Figure 2. The architecture of the proposed model along with the input/output dimensions of each layer.

The dataset is originally divided into TRAIN and TEST subsets. It has different speakers in the training and test sets. The training set consists of 462 speakers (3696 utterances), and the test set contains 168 different speakers (1,344 utterances). This split ensures that the models trained on the training set do not encounter the same speakers in the test set.

Since the original speech signals are clean, ground truth labels for VAD are derived using an energy-based method. It was observed that the obtained labels for the original speech data are unbalanced (most of the frames are active speech frames). This can pose problems in training the model. To balance the labels in the training set, silent sections are inserted into the original speech signals at random positions. Here, we insert a total of 8000 zero values which, based on the sampling rate, corresponds to approximately half a second of silence.

Moreover, to improve model robustness in real-world conditions, we augment the training data by adding noise. In order to simulate unpredictable real-world scenarios, random noise from the NoiseX-92 corpus is added to the clean speech signals [34]. NoiseX-92 contains various types of real-world noises, such as: White noise, Babble noise, volvo noise, Factory noise, Buccaneer noise, leopard noise. For augmenting training samples, for each original speech corpus, 2 types of noise between these noises have been randomly selected and for each noise, a random SNR belonging to the range of -5dB to 30dB has been randomly picked. In other words, for every speech sample, 2 new samples have been created by adding noise to it. This leads to create a diverse set of noise conditions.

2.2. MRMFCC Feature Extraction

In this research, Multi-resolution Mel-Frequency Cepstral Coefficients are extracted from the speech signal. The concept of multi-resolution, previously applied to cochleogram features in [35, 36], is employed here for the MFCC feature. This approach captures the spectral characteristics of speech signals at different resolutions. Specifically, three different window lengths are used to extract MFCCs from the speech signal.

The primary advantage of using Multi-resolution MFCCs is their ability to capture both fine-grained and coarse-grained spectral information. Traditional single-resolution MFCCs may not fully capture the variability of the speech signal, especially in noisy environments. Multi-resolution MFCCs extract features at multiple time scales: shorter windows capture fine-grained spectral variations, while longer windows capture more smoothed, coarse patterns. This enables the system to detect both rapid changes and more stable, long-term characteristics in the signal, resulting in a more comprehensive and noise-resilient feature representation.

Here, highest resolution features are obtained by extracting MFCCs using the smallest window length (25ms) and its hop size (5ms). Subsequently, the window length is doubled for the medium resolution, and doubled again for the lowest resolution. Here, the starting points of the two low resolution windows are aligned by the starting points of the high resolution windows. The MFCCs extracted at each resolution are concatenated to form a 1D vector for each high resolution frame, yielding a rich and detailed spectral representation.

To capture dynamic changes in the spectral features, the first- and second-order temporal derivatives, known as delta MRMFCCs and delta-delta MRMFCCs, are computed respectively. Utilizing these two features is common in speech processing applications, especially in VAD in which the distinction between speech and non-speech often depends on how the spectral features change [37, 38]. These derivatives help the model to detect the presence of speech even in situations where spectral features alone are not sufficient.

The obtained derivatives, along with the original MRMFCCs, form a three-dimensional representation for each frame. Considering the whole speech signal, the input F×3P×3 image is obtained, where F is the frame number, and P is the number of MFCCs extracted from each window (frame).

An example of a three-second speech signal (300 overlapping frames) along with its MRMFCC image is shown in Figure 1. In this figure, the P is considered 40. Each channel is displayed separately in figures (1.b), (1.c), and (1.d). In each channel, bands 1 to 40, 41 to 80, and 81 to 120 are dedicated to high resolution, medium resolution, and low resolution coefficients, respectively. Including both fine- and coarse-grained spectral details can help enhance the robustness and accuracy of the voice activity detection model in diverse acoustic environments.

2.3. Model Architecture

The proposed model, illustrated in Figure 2, incorporates convolutional and self-attention layers, using MRMFCCs as an input image with dimensions F×3P×3. In this context, F represents the frame number, and 3P results from concatenating P MFCCs extracted from each three resolution window. The three channels correspond to the MRMFCCs, delta MRMFCCs, and delta-delta MRMFCCs.

The proposed model, shown in Figure 2, includes a series of four convolutional layers, with the first three followed by an average pooling layer. For the convolutional layers, the filter number and kernel size are set to 32 and 3×3, respectively. To reduce computational complexity, the pool size of the pooling layers is set to 1×2. This reduces the dimensionality of the feature maps along the MRMFCC axis while preserving the temporal dimension. The output of the last convolutional layer is reshaped to a dimension of F×(32.3P/8), preserving the temporal dimension and flattening the feature maps.

To capture long-range dependencies and temporal relationships within the extracted features, two self-attention layers are utilized. Here, multi-head self-attention layers have been employed. This approach allows the model to focus on different parts of the sequence simultaneously. The mathematical description of the multi-head self-attention mechanism is as follows [39].

Given the input sequence $X \in R^{F*D}$ (here, F is the frame number, $D = 32.\frac{3P}{8}$), queries Q , keys K , and values V are obtained using linear transformations as:

$$Q = XW_Q , K = XW_K , V = XW_V \tag{1}$$

In which, $W_Q, W_K, W_V \in R^{D\times d_k}$ are weight matrices, and d_k is the dimension of the queries and keys.

For each attention head, the attention weight is obtained by applying softmax function on the scaled dot-product of the queries and the keys as [39]:

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

The output of the multi-head attention is obtained by concatenating the outputs of *h* head as:

$$MultiHead(Q, K, V) = concat(head_1, ..., head_h)W_O \tag{3}$$

Where each $head_i = Attention(Q_i, K_i, V_i)$, and $W_O \in R^{hd_k \times D}$ is an output weight matrix.

Before and after the multi-head attention layers, layer normalization is utilized to stabilize and normalize the inputs using the following equation:

$$LayerNorm(x) = \frac{x-\mu}{\sigma+\varepsilon}.\gamma+\beta \tag{4}$$

Where, μ and σ denote the mean and standard deviation of the input, and γ and β are learnable parameters. In the proposed model two multi-head self-attention layers are used, each with 4 heads $(h = 4)$ and a key dimension of 64 $(d_k = 64)$. With this approach, the proposed model can capture diverse patterns and relationships across the temporal dimension of the input sequence. The model ended by a dense layer with a sigmoid activation function. Each node is a probability value indicating the presence of speech activity. Notably, the weights of the dense layer are shared across each frame, ensuring that the model treats each frame consistently.

Using self-attention increases the model's complexity. As can be seen in figure 2, one self-attention layer includes approximately 492,000 trainable parameters (984,000 for the two layers used).

Table 1. Comparative analysis of accuracy across various methods.

SNR(dB)	P. M.			P.M without attention layers	GMM Clustering [11]	CNN-LSTM[26]	ACAM-LSTM[40]
	Precision%	Recall%	Accuracy%	Accuracy%	Accuracy%	Accuracy%	Accuracy%
-10	73.66	68.41	82.73	78.1	--	78.81	78.93
-5	80.63	85.21	88.18	83.6	63.4	79.87	83.46
0	90.29	93.89	93.74	89.1	77.8	85.51	88.59
5	94.01	95.34	95.97	91.1	84.4	88.74	90.82
10	94.73	95.58	96.56	91.5	92.6	89.18	91.13
15	96.23	96.66	97.35	93.8	95.42	90.79	93.01
20	97.23	96.97	97.85	95.0	97.3	91.49	94.24
25	97.81	97.05	98.07	95.8	--	93.86	95.63
30	97.87	97.24	98.16	96.1	--	94.85	96.37

To manage this complexity while maintaining proper accuracy, we used the minimum number of filters (32) in the CNN layers and incorporated pooling layers to reduce the size of the feature map. This lightweight CNN design allows for the inclusion of self-attention without significantly increasing the model's computational cost.

3. Experimental Results

To train the model, as mentioned in Section 2.1, training samples are generated from the TIMIT/TRAIN data. Briefly, the data generation process includes adding silence at random positions in the clean signal, extracting the VAD label with the help of an energy-based method from the clean signal, and adding noise to the signal (random type and SNR). By repeating this operation three times, adequate samples are generated to train the model. Moreover, the length of the training and test samples is considered to be three seconds. Accordingly, signals with a longer length are truncated, and in signals with a shorter length, silence is added at the end of the clean signal (before adding noise). MRMFCC images of each signal are created by considering windows of 25 milliseconds length and 10 milliseconds hop size. According to the signal length, window size, sampling rate, and number of MFCCs, the size of the obtained images is 300×120×3.

The proposed model is trained in 15 epochs using Adam's optimization algorithm with the learning rate of 0.0001. Binary cross-entropy is used as the loss function. All processes are conducted on a laptop with an Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz processor, 16 GB of RAM, and a GeForce GTX 1070 GPU with 8 GB of memory.

In figure 3, the graphs of loss and accuracy of the train and validation sets are shown over the epochs. It is worth noting that the validation set contains 200 test samples. As seen in the figure, the loss decreases over time. After 15 epochs, the validation error slightly becomes higher than the training error, suggesting that 15 epochs of training may be sufficient.

According to the number of original samples and the augmented ones (13800 samples in total), the training time is 13 minutes per epoch with a batch size of 50. By considering 15 epochs for training, the process of training takes approximately 3 and a quarter hours.

The evaluation of the proposed VAD is conducted using the TIMIT\TEST data. The speech signals are clean so an energy-based VAD is employed to generate the ground truth labels. To evaluate the robustness of the proposed framework, each speech signal is contaminated by an additive noise corpus. The noise type is randomly specified from the NoiseX-92 corpus, which consists of 15 noise types such as babble, Volvo, pink, etc. The experiments are conducted across various Signal-to-Noise Ratios (SNRs) from the set {-10, -5, 0, 5, 10, 15, 20, 25, 30} dB. Figure 4 shows an example of the experiment. Figure 4.a shows the original signal. Figures 4.b-4.d show three noisy signals which are obtained by adding the original signal with three different types of noise (babble, destroyer operation, and factory) at different SNRs. For each noisy signal, the ground truth VAD, the output from the proposed VAD model, and the VAD result after applying thresholding are depicted. As can be seen, the results demonstrate the robustness of the proposed framework under various noisy conditions.

The precision, recall, and accuracy are computed to evaluate the proposed VAD under varying noise conditions for each SNR level. These metrics provide a comprehensive view of the model's ability to correctly discriminant speech and non-speech segments.

These criteria are computed using following equations:

$$Precision = \frac{True \quad Positives}{True \quad Positives \; + \; False \quad Positives} \tag{5}$$

$$recall = \frac{True \quad Positives}{True \quad Positives \; + \; False \quad Negatives} \tag{6}$$

$$accuracy = \frac{True \quad Positives \; + \; True \quad negatives}{Number \quad of \quad Detection} \tag{7}$$

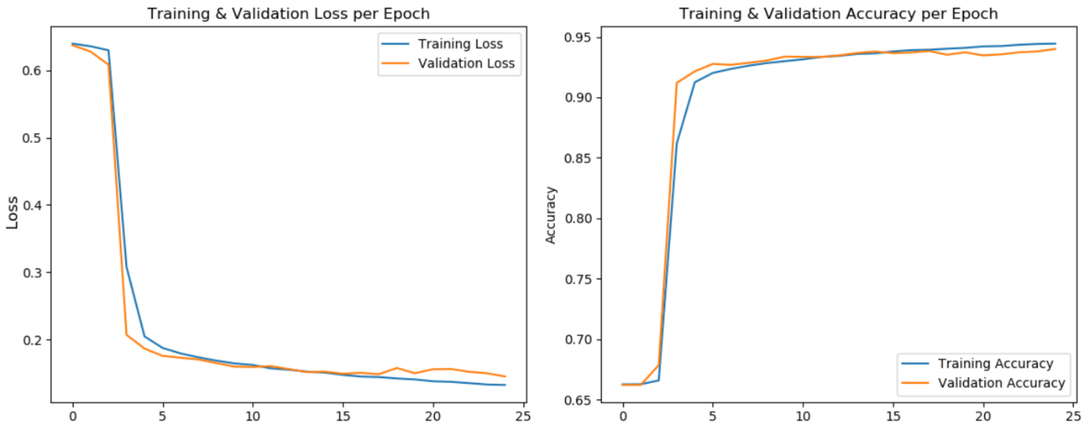


Figure 3. Training and validation loss and accuracy over epochs.

After performing the tests in each SNR, the mentioned criteria are calculated according to the proposed method and reported in Table 1. As can be seen, at lower SNR levels, such as -10 dB and -5 dB, the precision and recall values indicate the model's capability to maintain a reasonable balance between false positives and false negatives, even in noisy environments.

As the SNR increases, the model's performance improves significantly, with precision reaching 97.87% and recall 97.24% at 30 dB, demonstrating its robustness and high accuracy (98.16%) under clearer conditions. These metrics highlight the model's ability to detect speech effectively across a wide range of noise conditions, showcasing its adaptability and strong generalization to real-world scenarios where noise is variable.

Also, the accuracy of the proposed method has been compared with the approaches presented in [26,40,11].

- In [26], the proposed method is based on CNN-LSTM-dense architecture fed with the raw waveform. It begins with a frequency convolution layer followed by non-overlapping max pooling along the frequency axis. The output is then passed through several LSTM layers and one fully connected DNN layer before predicting two output targets.
- In [40], an adaptive contextual attention model (ACAM) along with LSTM-based attention approach has been utilized for voice activity detection. It employs a recurrent attention model that processes multiple input frames, focusing on the most crucial ones for classification. Its architecture consists of a decoder, attention mechanism, encoder, core network, and classifier.
- In [11], the proposed method utilizes clustering in the spectro-temporal domain. By applying Gaussian mixture models and WK-means clustering, the method reduces dimensionality and uses cluster attributes and energy levels to effectively distinguish between speech and silence.

In these three methods, TIMIT database has been used for evaluation. The accuracies of these method in the presence of noises similar to the noises used in this paper has been reported at different SNR levels and is shown in Table 1 along with the accuracy of the proposed method. In addition, to check the effect of attention layers, we have allocated a separate column in the table under the title of the "proposed method without attention layers".

Table 1 indicates that although the three selected methods are relatively robust in noisy conditions, the proposed method demonstrates superior accuracy. Also, the use of attention layers could improve the accuracy by almost 4-5%.

Finally, to further assess the effectiveness of the proposed method under varying noise levels, a final experiment was conducted by grouping the tests into three SNR ranges: -10 to 0 dB, 0 to 10 dB, and 10 to 20 dB. Precision-recall curves were plotted for each range instead of individual SNR levels to avoid overcrowding in the graphs, providing a clearer and more interpretable representation of the model's performance across these SNR intervals. This approach allows us to evaluate the robustness of the method effectively across distinct noise environments. Figure 5 presents the precision-recall curves and confusion matrices across different SNR levels. As shown in this figure, the proposed method demonstrates resilience to noise by maintaining high accuracy with low false positive (FP) and false negative (FN) rates.

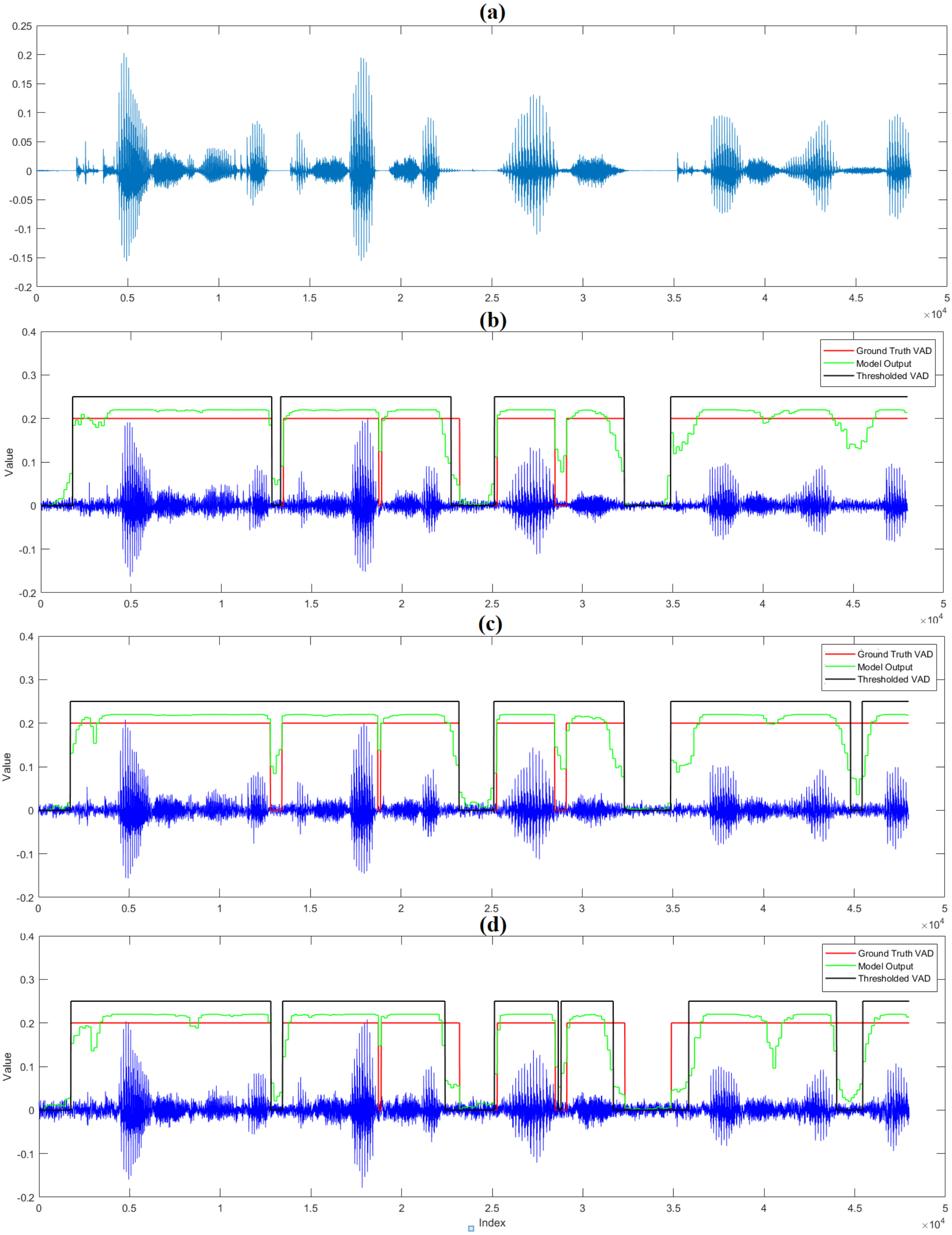


Figure 4. (a) Original speech signal. (b), (c), And (d) are noisy versions of (a), obtained by adding it with babble noise (SNR=7.6 dB), destroyer operations noise (SNR=7.9 dB), and factory noise (SNR= 5.6 dB), respectively. For each noisy signal, the corresponding ground truth VAD, model output, and thresholded VAD result are depicted. For better intuition and to avoid overlapping lines, the values of the ground truth VAD, model output, and thresholded VAD result have been scaled by 0.2, 0.22, and 0.25, respectively.

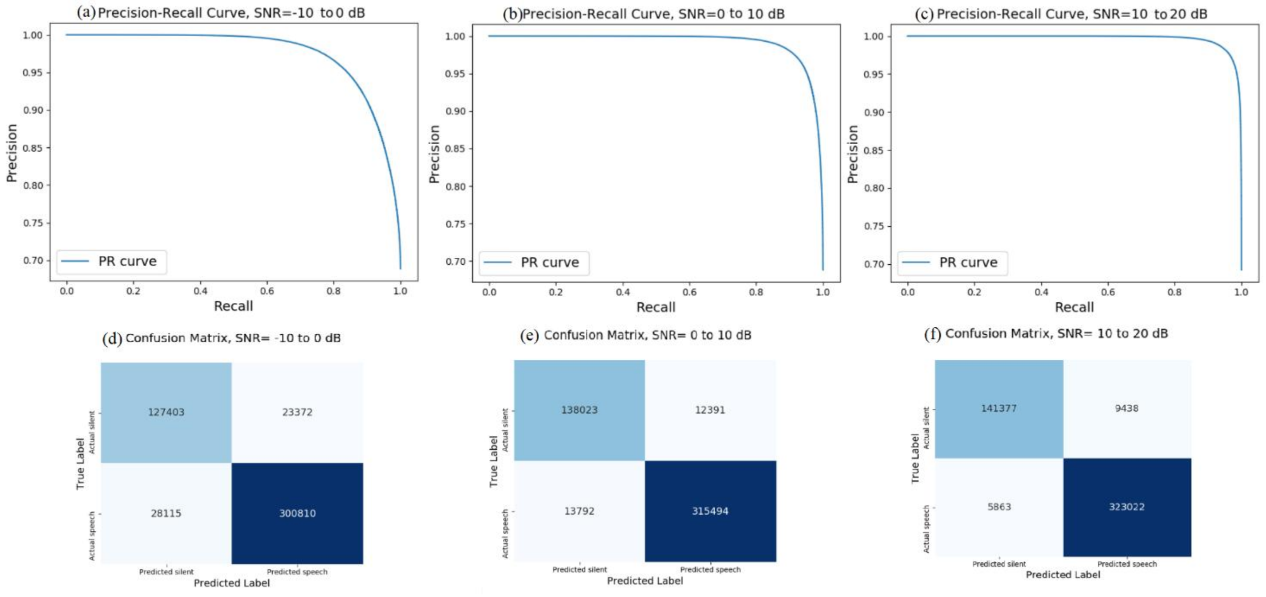


Figure 5. Precision-recall curves and confusion matrices across three SNR ranges (-10 to 0 dB, 0 to 10 dB, and 10 to 20 dB).

4. Conclusion

In this paper, a VAD method based on deep learning is proposed. In this method, a three-dimensional image of the audio signal is generated by concatenating MFCCs at three resolutions and considering their first and second derivatives in time. The proposed deep model consists of four convolutional layers, followed by two attention layers, and finally a dense layer. The training samples are created by adding noise to the clean data from the TIMIT/TRAIN dataset. For testing, the clean corpus, belonging to TMIT/TEST is combined with noise, and the model's accuracy is calculated for each SNR value. The accuracy of the proposed method is then compared with two recent state-of-the-art methods. The results demonstrate an improvement in accuracy of the proposed method compared to the other two methods. However, the model may face challenges in handling low-energy speech segments or speech under extreme noise conditions, where the distinction between speech and non-speech becomes more ambiguous. Future work could focus on improving the model's performance by exploring alternative feature extraction methods like wavelet transforms, evaluating other model architectures and, finally, optimizing the model for real-time deployment on low-power devices.

References

[1] M. W. Mak, & H. B. Yu, “A study of voice activity detection techniques for NIST speaker recognition evaluations”, *Computer Speech & Language*, vol. 28, no. 1, 295-313, 2014.

[2] Woo, K. Ho, T. Yang, K. Park, and C. Lee. "Robust voice activity detection algorithm for estimating noise spectrum." *Electronics Letters* 36, no. 2,180-181, 2000.

[3] T. H. Zaw, and N. War, “The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection”, *In 2017 20th International Conference of Computer and Information Technology (ICCIT) (pp. 1-5). IEEE*, 2017, December.

[4] Y. Kida, T. Kawahara, “Voice activity detection based on optimally weighted combination of multiple features”, *In INTERSPEECH* , pp. 2621-2624, 2005, September.

[5] F. Tao, & C. Busso, “Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection”, *In INTERSPEECH* (pp. 1938-1942), 2017, September.

[6] X.L. Zhang, & D. Wang, “Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection”, *In Fifteenth annual conference of the international speech communication association*, 2014.

[7] S. H. Chen, R. C. Guido, T. K. Truong, & Y. Chang, “Improved voice activity detection algorithm using wavelet and support vector machine”, *Computer Speech & Language*, vol. 24, no. 3, 531-543, 2010.

[8] S. M. Joseph, & A. P. Babu, “Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding”, *International Journal of Speech Technology*, 19, 537-550, 2016.

[9] D. Ying, Y. Yan, J. Dang, & F. K. Soong, “Voice activity detection based on an unsupervised learning framework”. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, 2624-2633, 2011.

[10] Z. Shen, J. Wei, W. Lu, & J. Dang, “Voice activity detection based on sequential Gaussian mixture model with maximum likelihood criterion”, *In 2016 10th*

International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 1-5). IEEE, 2016.

[11] N. Esfandian, F. Jahani Bahnamiri, & S. Mavaddati, “Voice activity detection using clustering-based method in Spectro-Temporal features space”, *Journal of AI and Data Mining*, vol. 10, no. 3, pp. 401-409, 2022.

[12] H. Veisi, & H. Sameti, “Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement”, *IET signal processing*, vol. 6, no. 1, pp. 54-63, 2012.

[13] X. Liu, Y. Liang, Y. Lou, H. Li, and B. Shan, ”Noise-robust voice activity detector based on hidden semi-markov models”, *In 2010 20th International Conference on Pattern Recognition* (pp. 81-84). IEEE, 2010.

[14] B. Liu, J. Tao, F. Mo, Y. Li, Z. Wen, & S. Liu,” Efficient voice activity detection algorithm based on sub-band temporal envelope and sub-band long-term signal variability”, *In The 9th International Symposium on Chinese Spoken Language Processing* (pp. 531-535). IEEE, 2014.

[15] N. Ryant, M. Liberman, & J. Yuan,”Speech activity detection on youtube using deep neural networks”, *In INTERSPEECH* (pp. 728-731), 2013.

[16] Y. Jung, Y. Kim, H. Lim, & H. Kim, “Linear-scale filterbank for deep neural network-based voice activity detection”, *In 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment* (O-COCOSDA) (pp. 1-5). IEEE, 2017.

[17] Y. Jung, Y. Choi, & H. Kim, “Self-adaptive soft voice activity detection using deep neural networks for robust speaker verification”, *In 2019 IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU) (pp. 365-372). IEEE, 2019.

[18] A. Sehgal, & N. Kehtarnavaz, “A convolutional neural network smartphone app for real-time voice activity detection”, *IEEE access*, 6, 9017-9026, 2018.

[19] M. H. Faridh, & U. S. Zulpratita, “HiVAD: A Voice Activity Detection Application Based on Deep Learning.”,*ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 9, no. 4, 856, 2021.

[20] P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, & F. Piazza, “Convolutional neural networks with 3-d kernels for voice activity detection in a multiroom environment”, *Multidisciplinary Approaches to Neural Computing*, pp. 161-170, 2018.

[21] P. Vecchiotti, E. Principi, S. Squartini, & F. Piazza, “ Deep neural networks for joint voice activity detection and speaker localization”, *In 2018 26th European Signal Processing Conference (EUSIPCO)* (pp. 1567-1571). IEEE, 2018.

[22] S. Mihalache, & D. Burileanu, “Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection”, *Sensors*, vol. 22, no. 3, 1228, 2022.

[23] R. Lin, C. Costello, C. Jankowski, & V. Mruthyunjaya, “Optimizing Voice Activity Detection for Noisy Conditions”, *In INTERSPEECH* (pp. 2030-2034), 2019.

[24] N. Wilkinson, & T. Niesler, “ A hybrid CNN-BiLSTM voice activity detector”, *In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6803-6807). IEEE, 2021.

[25] M. Ovaska, J. Kultanen, T. Autto, J. Uusnäkki, A. Kariluoto, J. Himmanen, & P. Abrahamsson” Deep Neural Network Voice Activity Detector for Downsampled Audio Data: An Experiment Report”. *arXiv preprint arXiv*:2108.05553, 2021.

[26] R. Zazo, T. N. Sainath, G. Simko, & C. Parada, “Feature Learning with Raw-Waveform CLDNNs for Voice Activity Detection”, *In Interspeech* (pp. 3668-3672), 2016.

[27] G. Gelly, J.L. & Gauvain, “ Optimization of RNN-based speech activity detection”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 646-656, 2017.

[28] J. Jia, P. Zhao, & D. Wang, “A Real-Time Voice Activity Detection Based On Lightweight Neural”. *arXiv preprint arXiv*:2405.16797, 2024.

[29] G. Dahy, A. Darwish,& A. E. Hassanein, Robust Voice Activity Detection Based on Feature Fusion and Recurrent Neural Network. In International Conference on Advanced Intelligent Systems and Informatics (pp. 359-367). Springer, Cham, 2024.

[30] Y. Korkmaz, Y., & A. Boyacı, . Hybrid voice activity detection system based on LSTM and auditory speech features. Biomedical Signal Processing and Control, 80, 104408, 2023.

[31] A. Sofer, & S. E. Chazan, “CNN self-attention voice activity detector”, *arXiv preprint arXiv*: 2203.02944, 2022.

[32] J. Thienpondt, & K. Demuynck,” Speaker Embeddings With Weakly Supervised Voice Activity Detection For Efficient Speaker Diarization”, *arXiv preprint arXiv*:2405.09142, 2024.

[33] J. S. Garofolo et al., TIMIT Acoustic-Phonetic Continuous Speech Corpus, Philadelphia, PA, USA:The Linguistic Data Consortium, 1993.

[34] A. Varga H. J. M. Steeneken, 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;12.3(1993): pp. 247-251,1993.

[35] R. Zhang, P. H. Li, K. w. Liang, & P. C. Chang, Voice Activity Detection by Jo1i nt MRCG and MFCC

Features with Robustness Detection based GRU Networks. In 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (pp. 1-2). IEEE, 2021.

[36] K. Raut, S. Kulkarni, & A. Sawant, . Multimodal Spatio-Temporal Framework for Real-World Affect Recognition. International Journal of Intelligent Networks, 2024.

[37] S. Alimi, & O. Awodele, Voice activity detection: Fusion of time and frequency domain features with a svm classifier. Comput. Eng. Intell. Syst, vol. 13, no. 3, pp. 20-29, 2022.

[38] S. Dwijayanti, K. Yamamori, & M. Miyoshi, Enhancement of speech dynamics for voice activity detection using DNN. EURASIP Journal on Audio, Speech, and Music Processing, 2018, 1-15, 2018.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, & I. Polosukhin. "Attention is all you need", *Advances in neural information processing systems*, 30, 2017.

[40] J. Kim, & M. Hahn, "Voice activity detection using an adaptive context attention model", *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181-1185, 2018.

رویکرد یادگیری عمیق برای تشخیص نواحی فعال گفتار مقاوم: ترکیب شبکه های کانولوشن و self-attention با استفاده از MFCC با وضوح چندگانه

خدیجه آقاجانی*

گروه مهندسی کامپیوتر، دانشکده فنی مهندسی، دانشگاه مازندران، بابلسر، ایران.

ارسال ۲۰۲۴/۰۷/۲۷؛ بازنگری ۲۰۲۴/۱۰/۰۲؛ پذیرش ۲۰۲۴/۱۱/۰۲

چکیده:

تشخیص نواحی فعال گفتار (VAD) نقش حیاتی در برنامه های مختلف پردازش صدا مانند تشخیص گفتار، تقویت گفتار، مخابرات، تلفن ماهواره ای و کاهش نویز دارد. عملکرد این سیستم ها را می توان با استفاده از روش VAD دقیق افزایش داد. در این مقاله، ضرایب کپسترال با فرکانس مل با وضوح چندگانه (MRMFCCs)، مشتقات مرتبه اول و دوم آنها (دلتا و دلتا۲)، از سیگنال گفتار استخراج شده و به یک مدل عمیق وارد می‌شوند. مدل پیشنهادی با لایه‌های کانولوشنی آغاز می‌شود. این ساختار در استخراج ویژگی‌ها و الگوهای محلی در داده‌ها مؤثر هستند. ویژگی های استخراج شده به دو لایه متوالی خود توجه چند سر وارد می شوند. با کمک این دو لایه، مدل می‌تواند به طور انتخابی بر روی مرتبط‌ترین ویژگی‌ها در کل توالی ورودی تمرکز کند، بنابراین اثر نویز کاهش می یابد. ترکیبی از لایه‌های کانولوشنال و توجه به خود، مدل را قادر می‌سازد تا هم ویژگی های محلی و هم سراسری را در سیگنال گفتار مورد بررسی قرار دهد. مدل پیشنهادی نهایتا با یک لایه متراکم برای طبقه‌بندی به انتها می رسد. لایه توجه چند سره عملکرد تشخیص کلی را با افزایش توانایی مدل برای تمرکز بر ویژگی‌های مرتبط در نقشه ویژگی ورودی، بهبود می‌بخشد. برای ارزیابی مدل پیشنهادی، از ۱۵ نوع نویز مختلف از پیکره NoiseX-92 استفاده شده است. نتایج تجربی نشان می‌دهد که چارچوب پیشنهادی در مقایسه با تکنیک‌های سنتی VAD، حتی در محیط‌های پر سر و صدا، عملکرد بهتری دارد.

کلمات کلیدی: تشخیص نواحی فعال گفتار، مکانیسم توجه به خود، ضرایب کپسترال فرکانس مل با وضوح چندگانه، یادگیری عمیق.