Voice Activity Detection by Joint MRCG and MFCC Features with Robustness Detection based GRU Networks

Rong Zhang, Pin-Hsuan Li, Kai-Wen Liang, and Pao-Chi Chang
National Central University, Taoyuan, Taiwan
rzhang.vaplab;phli.vaplab@gmail.com, kwistron@gmail.com, pcchang@ce.ncu.edu.tw

***Abstract*—In this paper, we proposed a Voice activity detection (VAD) model based on recurrent neural network(RNN) with joint MRCG and MFCC features. The system consists of two layers of gated recurrent unit (GRU) and beat the traditional methods in accuracy in our experiments.**

I. Introduction

The classification of voice segment from the voiced segment, unvoiced segment and silence segment in a speech signal is achieved using a Voice Activity Detector (VAD). It is a front-end and contribution to a large number of areas, such as automatic speech recognition, speech enhancement and so on. Various methods, traditionally based on feature engineering, statistical signal processing.

Our voice activity detection (VAD) model is a recurrent neural network (RNN) based classifier model. Our model consists of two layers of gated recurrent unit (GRU). The remainder of this paper is organized as follows: Section II presents the details of the proposed based Voice Activity Detector. In Section III, we present the experimental results. Section IV describes the compare our proposed framework with others, and then Section V concludes our findings.

II. Methodology

The GRU-based VAD is a frame-based classifier. First, we will introduce the feature extraction method, and the proposed based GRU. Figure.1 shows our architecture, where the audio input, the features are captured by Mel-Frequency Cepstral Coefficients(MFCC)[1] and Multi-Resolution Cochlea-gram(MRCG)[2].

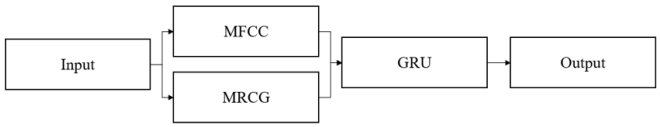

Figure 1. Subnetwork architecture of the proposed method

A. *Mel-Frequency Cepstral Coefficients*

In speech recognition and speaker recognition, the most commonly used speech feature is the Mel Inverse Spectral Coefficient, which is particularly suitable for speech recognition, considering the degree of perception of different frequencies by the human ear. MFCC extract the features by the following steps:

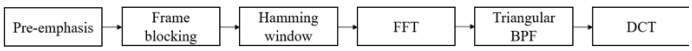

Figure 2. MFCC architecture

B. *Multi-Resolution Cochlea-gram*

The other way of extracting features is MRCG, it is a multi-resolution filter feature, which is expressed by combining multiple cochleae with different resolution.

The high-resolution cochlea-gram focuses on capturing the local information of the signal, the low-resolution cochlea-gram grasps the global information.

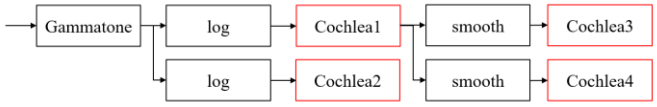

Figure 3. MRCG architecture

MFCC extract the features by the following steps:

I. Calculating the Cochlea-gram of the first channel, the frame length is selected as 25ms, frame shift is 10ms, Cochlea1 is obtained after logging all the time frequency units.

II. Cochlea2 is obtained after selecting 200ms frame length and 10ms frame shift, and taking logarithms of all the time frequency units.

III. Using 5*5 window to Cochlea1, smooth and average pool to get Cochlea3. Cochlea4 and Cochlea3 calculation is similar, using 11*11 window to Cochlea1, smooth and average pool to get Cochlea4.

IV. The MRCG is obtained by concatenating Cochlea1 to Cochlea4.

C. *Model Architecture*

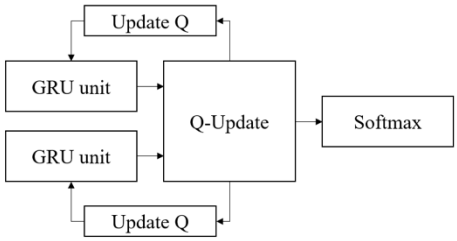

Figure 4. Model Architecture

The architecture uses Robustness Detection as an updating method. GRU is used for the core network. Hierarchical Q-Learning[3] is used for determine robustness between subnetworks. The algorithm is shown in Figure 5. In following algorithm, hierarchical Q-function is used due to reduce time complexity.

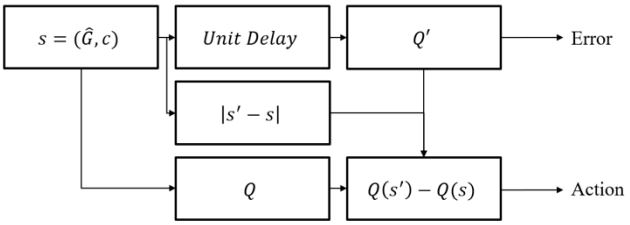

Figure 5. Hierarchical Q-Learning Architecture

And a decode FSM is used for reduce memory, which is shown in Figure 6.

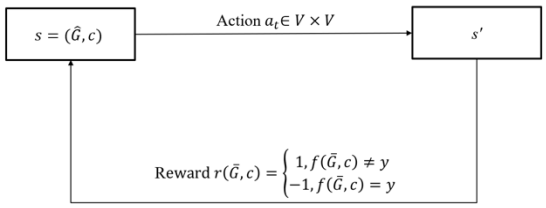

Figure 6. FSM Architecture

III. Experiments and Results

A. Dataset

Our training dataset is using musan[4], which is a corpus of music, speech, and noise contains about 60 hours of speech data. The dataset we use with a number of 400, a maximum length of 612 seconds, and a minimum length of 341 seconds, for a total length of 63.7 hours.

B. Experiments and results

The experimental results are shown in Figure 7 and Figure 8, and it can be seen that the original voice signal can be classified into three types, voiced, unvoiced, and silent. These three categories do not overlap with each other.

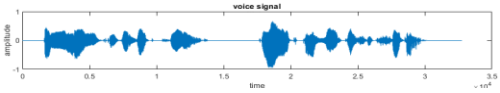

Figure 7. test1 signal from test set

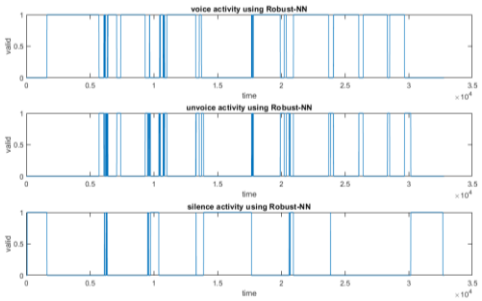

Figure 8. Based on the proposed, the experimental results of test1

IV. Evaluation

A. Robustness Detection

The blue line represents average accuracy of whole model, and the others denotes subnetworks. Ones return to zero means being eliminated, i.e. red and gray.

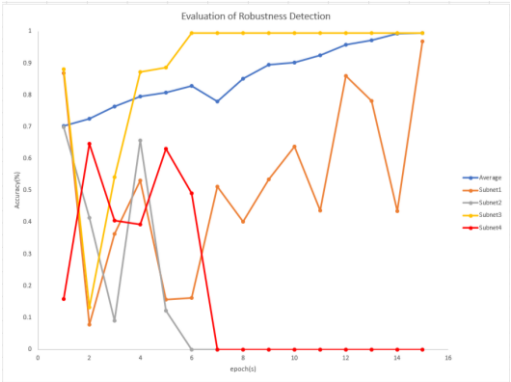

Figure 9. Robustness of subnetworks

B. Accuracy

The accuracy of our proposed method is 0.2% higher compared to the ACAM-based architecture, also improves the accuracy by nearly 9% over the SVM-based method.

Method	Accuracy
Ensemble-SVM	90.52
ACAM	99.2
Proposed	99.41

Table 1. Accuracy Compared to others.

V. Conclusion

In this paper we investigated the use of a neural network architecture in VAD, and using MFCC features and MRCG features. Our experiments show that GRU-based VAD outperforms other traditional methods and the addition of reinforcement learning has improved the accuracy.

Reference

[1] Md. Afzal Hossan, Sheeraz Memon, and Mark A Gregory, "A Novel Approach for MFCC Feature Extraction", *International Conference on Signal Processing and Communication Systems*, 2010.

[2] J. Chen, Y. Wang and D. Wang, "A feature study for classification-based speech separation at very low signal-to-noise ratio," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[3] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song, "Adversarial Attack on Graph Structured Data", arXiv:1806.02371v1 [cs.LG] 6 Jun 2018.

[4] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A Music, Speech, and Noise Corpus", *Language and Speech Processing*, 2015

[5] Juntae Kim and Minsoo Hahn, "Voice Activity Detection Using an Adaptive Context Attention Model", *The international Conference on Acoustics, Speech, & Signal Processing*, 2019

[6] Jayanta Dey , Md. Sanzid Bin Hossain , and Mohammad Ariful Haque, "An Ensemble SVM-based Approach for Voice Activity Detection", *Communications, Speech and Vision*, 2019