Master's Thesis

Study on Robust Voice Activity Detection Using CNN Encoder-decoder Based on MTF Concept Under Noisy Conditions

LI, Nan

Supervisor Masashi Unoki

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2020

Abstract

In recent years, emerging industries such as smart homes, dialog robots, and smart speakers have flourished, which has dramatically changed people's lifestyles and the way people interact with machines. Voice interactions have been widely used in these emerging fields. With the application of deep learning in speech task (e.g., automatic speech recognition (ASR), emotion recognition, and other speech classification tasks), the performance of these speech tasks has been greatly improved. The speech recognition accurate rate also has exceeded, and the recognition effect has basically reached the level of human hearing. The above tasks always rely heavily on voice activity detection (VAD) technology.

VAD is a technique to determine the beginning and end points from a segment of the speech signal. VAD is very important in front-end processing in many different speech applications. Effective VAD can: (1) eliminate the interference of the silent segment or noise segment for the speech task. (2) reduce the amount of computation for the computer. Based on the above two points, the useful VAD could improve the speed and accuracy of speech tasks, especially all the speech applications first step is VAD. There are many methods to study VAD in previous research. However, such methods remain the robustness and accuracy insufficient, the performance of previous methods is often affected by noise. Almost all the VAD methods can be used in clean or stationary noise environments, only a few of methods could be used under low signal-to-noise noise (SNR) non-stationary environments. It's easy to think of noise as speech, removing the effects of noise conditions is crucial for the speech or non-speech detection task.

To find out what features and underlying concepts can be used to solve noise issue and then to propose an accurate and robust VAD method against environmental conditions, this research by incorporating the modulation transfer function (MTF) concept into deep neural networks architecture aims to solve the above issue final improve the VAD accuracy under noise environments. The effect of noise on noisy speech can be regarded as MTF. In theory, if the MTF for the speech under noisy environments can be estimated, the impact of noise on speech could use MTF to reduce, and further obtain an accurate and robust VAD performance. In this research, a denoising method is used for the temporal power envelope restoration. By setting a threshold for the restored temporal envelope a further determine speech or non-speech (VAD) can be obtained. Therefore, the key to this research is to find a way to do temporal envelope denoising.

To eliminate the effect of additive noise, global signal-to-noise (gSNR) should be estimated. Previous work proposed a robust gSNR detection method in the sub-band speech signal. But with the increase of noise (decreased SNR), because VAD cannot usually be accurately judged by a single threshold in the whole utterance, this method is not robust at low SNR. This method is also often used in a specific kind of noise environment, and another environment requires changing parameters. Many people have proposed deep neural networks (DNNs) based end-to-end gSNR detection methods, end-to-end based methods usually extract acoustic features of one utterance and then input all of them to a neural network to predict gSNR. This method often has the problem of data mismatch or environment mismatch. In addition, the end-to-end method requires that the input utterance have the same shape. If they are different, you need to do pooling and speech cutting to make all sentences the same length. This kind of processing method will cause the problem of out of memory and cannot adapt to all applications.

In order to solve the above gSNR estimation problems, an indirect gSNR estimation method is proposed in this research. Because the additive noise and clean speech components are mixed together, estimating the gSNR in the original time domain signal is very difficult. In this research, the sub-band signal processing method is used. The proposed gSNR estimation method mainly includes sub-band speech signal processing, sub-band threshold calculation unit, sub-band power calculation unit and gSNR calculation unit. Motivated by the noise and speech that have a difference represent in the sub-band, the constant-band filter-bank (CBFB) is used to split noisy speech into a different sub-band speech signal. The sub-band-based processing method makes speech and noise processing more accurate than global full-band processing. In addition, in previous studies, a static threshold is often used to judge VAD, but this judgment is not reasonable because different speech sample may be in different threshold or different noise ratio. In this study, based on the sub-band and the convolutional neural network (CNN) encoder-decoder (C-ED) structure we propose a gSNR estimation method, this method could estimate the noise ratio of different sub-band speech signal sample. The deep neural network could solve the non-linear problems well, this research uses for C-ED framework to estimate sub-band speech signal threshold about speech and noise. Finally, the final gSNR can be obtained according to the obtained threshold.

Experiments conducted on the stationary and non-stationary conditions demonstrate that the proposed C-ED based MTF method achieves better performance compare with the previous MTF based VAD method. This method can effectively reduce the bad affect of noise for the VAD, especially in the environment of low SNR and non-stationary noise.

In conclusion, a VAD system utilizing the convolutional neural network encoder-decoder model has been proved to achieve better performance compared to the previous modulation transfer function based method.

A VAD system utilizing a C-ED model was proposed in this research. The proposed method can improve the accuracy of speech/non-speech detection similar to the previous method using the modulation transfer function. Although there was still much room for performance improvement, this research put the first step toward the realization of incorporating deep neural networks into modulation transfer function concept. Moreover, this will further provide key technical support for not only various speech applications but also man-machine speech communications under real environmental conditions.

Acknowledgment

First of all, I want to thank JAIST's Professor Massashi Unoki, who taught me how to organize logic and have a scientific idea. I really appreciate every suggestion he gave me, and thank him for his guidance on my research. It is impossible to complete this paper without his guidance and support.

In addition, I thank Professor Masato Akagi for his guidance and advice. His guidance has improved my research.

Thanks to the joint training program of Tianjin University and JAIST for allowing me to experience two different research models in China and Japan, which will be very helpful for my future research path.

At the same time, thank Professor Dang Jianwu and Professor Wang Longbiao for their support. Thank them for introducing me to JAIST, and for giving me the opportunity to study here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the research background, research issues and research objective are briefly introduced. First of all, we explain the concept of voice activity detection (VAD) and introduce its history. Secondly, we describe the issue statement of current studies. Thirdly, the research objective of our study is presented. Finally, the organization of this paper is listed.

1.1 Research background

In recent years, emerging industries such as smart homes, dialog robots, and smart speakers have flourished, which has dramatically changed people's lifestyles and the way people interact with machines. Voice interactions have been widely used in these emerging fields. With the application of deep learning in speech task (e.g., automatic speech recognition [1], speaker recognition [2], emotion recognition [3], speech enhancement [4, 5] and other speech classification tasks), the performance of these speech tasks has been greatly improved. The speech recognition accurate rate also has exceeded 95%, and the recognition effect has basically reached the level of human hearing. The above tasks always rely heavily on VAD technology.

The aim of VAD is to judge speech and non-speech boundaries in the speech signal. VAD is very important in front-end processing in many different speech applications, e.g., speech synthesis [8], speech recognition [7], speech enhancement [6], and speech classifications [9]. Effective VAD can: (1) eliminate the interference of the silent segment or noise segment for the speech task. (2) reduce the amount of computation for the computer. Based on the above two points, the useful VAD could improve the speed and accuracy of speech tasks, especially all the speech applications first step is VAD. Figure 1.1 is a flowchart often used in speech applications.

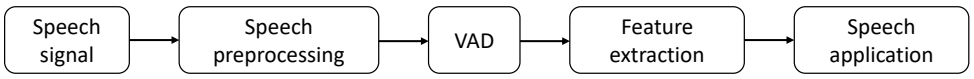

Figure 1.1: A flowchart often used in speech applications

Because of the importance of VAD, the history of its research can be traced back to the 1980s [10]. The initial VAD methods are mainly based on traditional signal processing methods, these methods mainly extract some acoustic features, and then set some simple thresholds to judge speech and non-speech. The above-selected acoustic features generally can easily determine the boundaries between speech and non-speech, conventional linear spectrum frequency, zero-crossing rate, signal energy, correlation function [11] wavelet [12] and pitch features [13] widely used in VAD methods based on signal processing.

There is many typical signal processing based VAD methods, e. g., signal energy-based method [11], VAD in G.729B [14], AMR VAD methods [15, 16], empirical mode decomposition (EMD) based VAD method [17, 18], and modulation spectrum analysis (MSA) based VAD method [19, 20]. Most of them have two-step processes: the first step is extracting acoustical features from the observed signal and the second step is classifying speech or non-speech. Although these features and classifications have excellent performance under clean conditions, their performances are drastically reduced by interference conditions.

With the application of statistics in various fields, e.g., image recognition [21], natural language processing [22], recommender system [23], and speech recognition [24], many scholars started using statistics to the VAD [25, 26, 27, 28, 29, 30] in the late 1990s. Statistical learning usually includes two categories: supervised methods and unsupervised methods.

Both supervised and unsupervised can be used for VAD, some people use supervised methods, and some people use unsupervised methods. Hidden Markov model (HMM) [26] was first used in VAD as a supervised method. The first step in this method is to extract acoustic features. Features that can better distinguish between speech and non-speech are often used, e.g., energy, spectrogram, and wavelet always be used in this step. Lastly, by using HMM to calculate the maximum posterior probability suitable for determining speech and non-speech we also could do VAD task. As for unsupervised methods, clustering [31] is firstly used. The clustering-based method also has the acoustic feature extracted and speech or non-speech to determine two steps. This method determines the speech or non-speech by counting the distribution of speech features. With the application of deep learning to

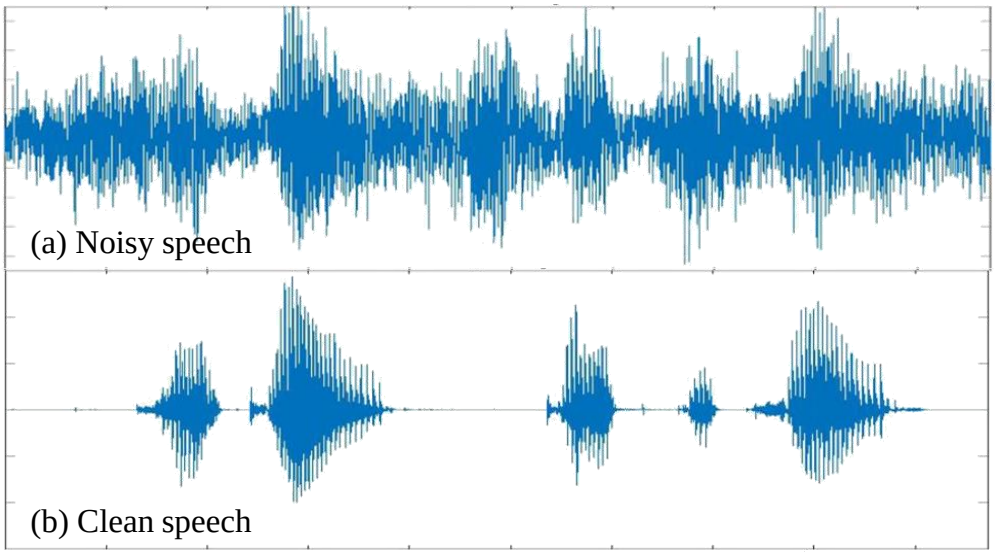

Figure 1.2: Example of noisy speech and clean speech

various fields, the VAD method based on deep learning greatly increases the accuracy of previous methods.

1.2 Research issue

Although there are a variety of VAD methods, almost all the current methods are used for clean and stable noise environments. But there are all kinds of noise in the real world, and these noises will seriously interfere with the recognition effect of speech tasks with the increase of noise, the recognition rate of automatic speech recognition will decrease rapidly. Figure 1.2 is a comparison of noise and speech, as we can see noisy speech always disorganized. The same with the automatic speech recognition, no matter what concept and feature are used in the VAD task if above VAD methods are used under noisy especially non-stationary noise (e.g., factory and babble) and very low signal-to-noise ratio (SNR) conditions, performance is drastically decreased. Non-stationary noise and very low SNR conditions also are problems that cannot be solved by industry and academia for the VAD task.

1.3 Research objective

The objective of this research is to propose a robust VAD method that even in the non-stationary and low SNR noise conditions also

could get a good performance.

To achieve these purposes, this study aims to find out what features and underlying concepts can use to solve the above issue and then propose an accurate and robust VAD method against environmental conditions. Since the effects of noise have complexly a bad influence on VAD, the recent research on VAD has to consider how to reduce these effects for detecting speech. Thus, the research purpose in this study is to propose a VAD method that could robustly and correctly detect speech and non-speech from the objective signal under noisy conditions even if noise could not be mathematically modeled. This study definitely contributes the advanced technical support for different speech applications, e.g., speech classification, speech recognition and speech enhancement in real environments. The proposed method can also contribute a key technique for speech communications.

1.4 Organization of this research

The rest of this article is as follows:

- Literature Review (Chapter 2): In this chapter, we will give an overview of voice activity detection, modulation transfer function and the receiver operating characteristic based global SNR.
- The previous method (Chapter 3): In this chapter, the previous modulation transfer function based VAD method is given a detail introduction.
- The proposed method (Chapter 4): We describe our modulation transfer function based VAD method for using the CNN encoder-decoder structure to calculation the global signal-to-noise ratio.
- Experimental and Results (Chapter 5): We give detail to the dataset used in this research. We also carried out some objective measurements to evaluate the correctness of the extracted sub-band signal and temporal power envelope. Then we present the results of the objective measurement and subjective test.
- Conclusion (Chapter 6): This chapter summarizes our work and we also give out the contribution of our method. We also give out the remaining work for our research, these remaining works are future research direction.

Chapter 2

Literature Review

As mentioned above, whether it is a statistical-based method or a signal-processing-based method, almost all VAD algorithms have two steps: feature extraction and speech and non-speech decisions. This chapter mainly introduces some typical VAD methods based on conventional signal processing and statistic. The basic principles of deep learning in detail and details the commonly used deep learning-based VAD methods also introduced in this chapter. This research also points out some problems in previous VAD methods, and we hope to solve these problems to achieve robust VAD performance through research.

2.1 Overview of voice activity detection

There are many kinds of VAD algorithms, and the most widely used methods are based on signal processing and statistics. Among them, the method based on signal processing mainly extracts some distinguishing speech features and then sets the corresponding threshold to judge VAD. The statistical-based detection method mainly extracts the characteristics suitable for distinguishing between speech and non-speech and then determines VAD according to the maximum posterior probability of speech or non-speech. But its adaptability to noise has been the bottleneck of research. The detection algorithm based on statistical characteristics can have a good suppression effect on specific noise, but its adaptability to different environments is not very good, and there will be many problems of environment mismatch and data mismatch. The traditional VAD algorithm flow in a noisy environment is shown in Figure 2.1.

As shown in Figure 2.1, the first step of the VAD system is speech preprocessing. To make the VAD system work well under noisy conditions, some

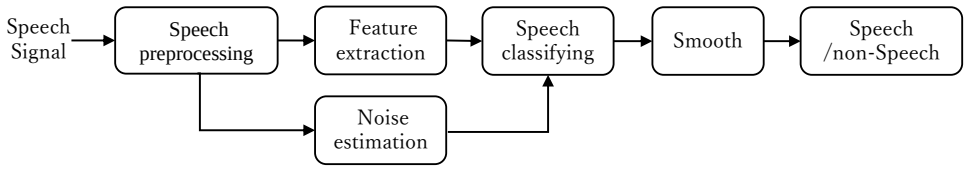

Figure 2.1: Overview of a VAD system

research adds a speech enhancement preprocessing. Since a speech utterance is time-varying, this makes the entire system difficult to process. In order to process voice signals more easily, a frame operation is often used to divide the speech into many small segments. We can think of these small pieces of speech as time-invariant. To reduce the bad affect of additive noise on VAD, the noise needs to be estimated.

The most common VAD system is mainly composed of feature extraction and speech/non-speech determination (classification). In the following section, we will display some typical method of these two stages.

2.1.1 Feature extraction

The feature extraction module in the VAD algorithm is mainly divided according to the characteristics of its speech. Among them, there are many time-domain features of the speech such as the short-term zero-crossing rate (ZC), autocorrelation, and logarithmic energy. There are also many frequency-domain features of the speech. It mainly includes signal features such as mel-frequency cepstrum coefficient (MFCC) features, spectral entropy, and long-term spectral differences. At the same time, it also has mixed features such as the combination of frequency domain and time domain, and wavelet domain features. When detecting endpoints of speech, these features play a decisive role in distinguishing between speech and noise. The extracted features are expected to have the following characteristics: (1) this feature must be easy to extract (2) this feature value must be able to effectively determine speech and non-speech, or the difference between them is stable and easy to distinguish (3) the voice characteristics expressed by this feature change with increasing noise not sensitive. Therefore, choosing proper feature values has a great impact on VAD, we also will give an overview of these features.

Short-time signal energy Speech signals are usually divided into unvoiced and voiced. Voiced voices have short-term periodicity and energy concentration. Unvoiced signals have noise-like characteristics, which are often mixed with noise and difficult to detect. Therefore, we can think that

the short-term energy of the noise frame is much smaller than that of the speech frame signal. The short-term energy calculation is as

$$E_n = \sum_{m=n}^{n+N-1} [y(m)w(n-m)]^2 \tag{2.1}$$

where w is window function of window length N. Energy-based features cannot adapt to noisy environments.

Short-term average zero-crossing rate (STAZC) The amplitude of the signal must pass through a zero value from the positive value to the negative value, and also pass through a zero value from the negative value to the positive value. The number of times a statistical signal crosses zero in one second is called ZC. The speech waveform is divided into a plurality of small speech waveform, and the zero-crossing rate of each segment of the speech signal is statistically averaged, which is called a STAZC.

STAZC can be used to judge unvoiced or voiced speech in speech signal analysis. It can be known from the speech generation model that when the voiced sound is generated, the vocal cords vibrate. Although there are several resonance peaks in the vocal tract, the glottal wave causes high-frequency fading of the spectrum, so the voiced energy is concentrated in the range of 3kHz. Conversely, in the unvoiced voice, the vocal cords do not vibrate, and some parts of the channel block the airflow to generate white-like noise, and its energy is concentrated in a higher frequency range. Since the low frequency corresponds to the low ZC, and the high frequency corresponds to the high ZC, there is a corresponding relationship between the ZC and the unvoiced and voiced sounds of the speech. The short-term average ZC is calculated as follows

$$\begin{aligned} z_n &= \sum_{m=-\infty}^{\infty} |\mathrm{sgn}[y(m)] - \mathrm{sgn}[y(m-1)]| \cdot w(n-m) \\ &= |\mathrm{sgn}[y(n)] - \mathrm{sgn}[y(n-1)]| * w(n) \end{aligned} \tag{2.2}$$

where $*$ is convolution calculation, $w(n)$ is window function.

The STAZC of the voiced sound is high, and STAZC of the voiced sound is low. However, there are overlapping areas between the two distributions, so it is not easy to obtain the ideal VAD based on the STAZC to accurately determine unvoiced and voiced sounds. Therefore, in practice, multiple features parameters of speech are often used.

Mel-frequency cestrum coefficient (MFCC) MFCC is widely used in many speech algorithms. It was originally conceived that the auditory characteristics of the human ear should be taken into account in the speech

features, and it has a non-linear correspondence relationship with the speech frequency, which makes the calculation accuracy of MFCC decrease as the frequency increases. Using this feature will discard the high-frequency information and select select the low-frequency information of the speech. Performing MFCC includes the following steps.

(1) **Pre-emphasis**: Speech signal passes a high-pass filter

$$H(Z) = 1 - \mu Z^{-1} \tag{2.3}$$

where μ is generally taken between 0.9-1.0 to improve high-frequency information.

(2) **Framing** Because the audio signal is non-stationary, but many audio processing technologies are based on a probability model, there is a requirement for the signal: the signal is a stationary signal. Otherwise, statistics such as mean and variance are meaningless. Usually the audio signal is framed to solve this problem. It is assumed that each frame is stable. Generally, 20-30ms is used as a frame, with an overlap rate of 25%, 50%, and 75%.

(3) **Window function** In order to avoid spectrum leakage, a Hamming window is often used for processing.

(4) **Mel-filter bank** After the speech waveform passes through the Mel filter bank, the conversion relationship between the linear frequency and the Mel frequency is

$$f_{mel} = 2595 \cdot \log\left(\frac{f}{7000} + 1\right) \tag{2.4}$$

On the Mel-axis, P is equally divided, and the power spectrum is added according to the triangular window on the Mel-axis, and the Mel sub-band energy $M_1, M_2, ...M_n$ can be obtained.

(5) **Discrete cosine transforms** After the information of each frequency band is separated, Discrete cosine transforms (DCT) calculation can be used to obtain the final features, the following function is DCT

$$D_k = 2\sum_{n=0}^{p-1} M_{n+1} \cos\frac{2n+1}{2P} k\pi \tag{2.5}$$

Since the MFCC feature takes into account the auditory characteristics of the human ear and does not have any assumptions, this feature has good speech recognition performance and noise immunity.

2.1.2 Speech/non-speech decision

(1) **Threshold** The threshold-based speech/non-speech determination method is the simplest VAD method, and its voice and non-voice judgment methods

mainly compare the extracted voice features with a set threshold and judge VAD through preset rules.

If the thresholds are set properly, the algorithm can detect speech and flying voices better. However, since the decision thresholds need to be set based on experience, the thresholds play a important role in the entire VAD task. Therefore, how to set an adaptive, accurate and reliable threshold according to the additive noise environment is still a issue to be solved.

(2) **Gaussian mixture model (GMM)** The basic principle of the VAD algorithm based on GMM [32] is to use for the features for each frame of speech and noise waveform, and divide these features into several classes, assuming that classes are independent of each other, and between classes and vectors. Then the vectors in each class are the same feature distribution, and the normal distributions of multiple classes are added by a certain weight to obtain the overall distribution of the speech and noise feature vectors. Next, a speech model and a noise model are established based on the training average, covariance, and threshold parameters. The input signal of each frame is determined by the principle of maximum posterior probability to determine whether it is speech or noise, and the model parameters are updated appropriately. However, due to the problem of data mismatch and environment mismatch in real life, this method does not produce good results in real life.

2.1.3 Smooth processing

In VAD detection, decision smoothing is a very important link. Its quality directly affects the final detection result, because in most VAD detection, the signal is framed by the frame, and then the endpoint is judged by frame. In the process, we often encounter the problem of speech clipping, so that the VAD algorithm must meet better robustness in a noisy environment. Therefore, the VAD algorithm generally needs to add a decision smoothing module. At the same time, we need to pay attention to the following principles. First, because speech recognition and other algorithms have higher requirements for the starting point of the speech, we generally need to push the smoothing point forward for 0.2 seconds, and the cutting problem in speech can be resolved by speech. The correlation between the length of the silent sound and the corresponding number of frames is delayed accordingly, and the problem existing at the end of the speech sentence can only be solved by artificial settings. As a result, the speech at the end of the sentence is confused with noise, which causes some errors. In addition, since the transient noise is relatively close in time to the speech segment of the speech, it is difficult to distinguish them, and a better algorithm needs to be studied to solve this problem.

2.2 Deep neural network in voice activity detection

2.2.1 Artificial neural network (ANN)

ANN [33] is a computational model that simulates the processing of information by the human brain. In the past decade or so, with the continuous deepening of the research work related to ANN, great progress has been made. ANN has solved many practical problems which are difficult to be solved by modern computers in many fields (e.g., pattern recognition and automatic control).

In artificial neural networks, the smallest computing unit is a neuron. It receives input parameters from other neurons and outputs the final result after calculation. As shown in Figure 2.2, each input of neurons many weights (w), and the input parameter also contains a very important parameter bias (b). And each neuron will apply the following function to get the final output t

$$t = f\left(\overrightarrow{\mathbf{W}}\mathbf{A} + b\right) \tag{2.6}$$

where $\mathbf{A} = [a_1, a_2, ..., a_n]$ are inputs of neuron, $\mathbf{W} = [w_1, w_2, ..., w_n]$ is the weight of each synapse of the neuron. f is the activation function, which is usually a non-linear function. In real life, many things are not linear curves. Therefore, in order to better fit the calculation laws in life, non-linear curves are needed to fit these real laws. This non-linear fitting process is called activation, and this non-linear function is called activation function. Common activation functions are as follows:

Sigmoid After the Sigmoid function, the result is a natural number in

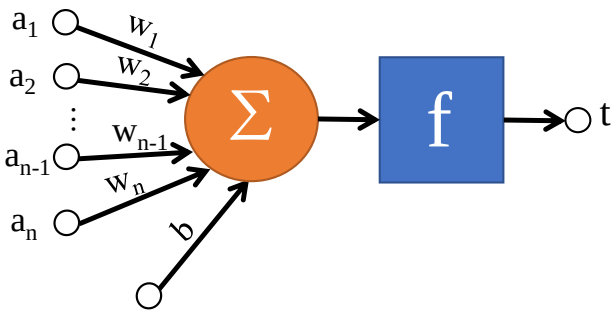

Figure 2.2: Single neuron

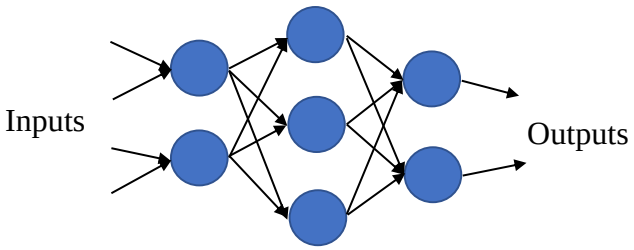

Figure 2.3: Multilayer perceptron

the interval [0,1].

$$f(x) = \frac{1}{e^{-x} + 1} \tag{2.7}$$

Tanh The result obtained after Tanh is a value between [-1,1].

$$f(x) = -\frac{e^{-x} - e^{x}}{e^{-x} + e^{x}} \tag{2.8}$$

ReLU After ReLU, the result is a number between $[0, \infty]$.

$$f(x) = \max(0, x) \tag{2.9}$$

In addition, a neuron cannot represent complex operations in real life. In order to achieve more complex operations, we will use multiple layers of network and multiple neurons to form a multilayer perceptron [34]. As shown in Figure 2.3, this is the simplest three layers multilayer perceptron which the last layer is the output layer, the begin layer is the input layer, and the middle layer is hidden layer.

2.2.2 Deep neural network-based voice activity detection

The traditional VAD method performs poorly under the condition of the low signal-to-noise ratio. As neural networks are gradually applied to speech and image tasks, research has begun to use deep neural networks to perform VAD tasks to the improvement of VAD performance at low SNR. A preferred choice for VAD is a deep neural network (DNN) [35], as shown in Figure 2.4, which has been extensively explored in the past few years. First, a DNN classification model is trained from a set of noisy speech represented by multiple speech features (e.g., Pitch, MFCC, LPC, and PLP). Then, the features of noisy speech are added to the trained DNN model to generate the maximum posterior probability of speech and non-speech. As

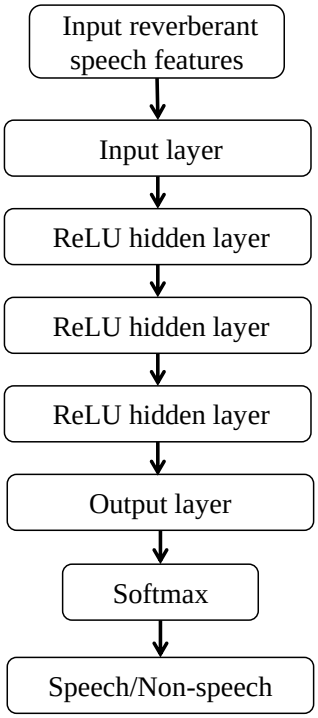

Figure 2.4: DNN based VAD method

with other classification tasks, the loss function used by deep learning-based VAD methods is cross-entropy (CE) which is calculated from the following function

$$\textbf{Loss} = -[(1-y)\cdot\log(1-p) + y\cdot\log(p)] \tag{2.10}$$

where y represent the ideal state, it is the true speech or non-speech. p is the probability that the sample is predicted to be positive.

In addition to DNN [35, 36], convolutional neural networks (CNN) and long short-term memory (LSTM) [38] will also promote the robustness of VAD. But DNNs-based VAD methods are also very restrictive, it must have massive datasets including these noise conditions, for the unknown environments need retraining model.

Chapter 3

Previous method

Morita et al. proposed a robust VAD method using the modulation transfer function (MTF) based concept in previous studies [20]. We will give a brief introduction to MTF-based method in this chapter. In addition, the key global signal-to-speech (gSNR) estimation methods proposed in the previous methods [39] will also be described in detail in this research.

3.1 Modulation transfer function based VAD

3.1.1 Modulation transfer function concept

The definition of MTF was proposed in the speech intelligibility prediction task in room acoustics by Houtgast and Steeneken [40]. It is used as a modulation index that illustrates the relationship between the degree of modulation of the temporal envelope between the input and output signals in the enclosure. The input temporal power envelope and output temporal power envelope could define as follow function

$$\textbf{Input} : \overline{I_i^2}(1 + \cos(2\pi f_m t)) \tag{3.1}$$

$$\textbf{Output} : \overline{I_o^2}(1 + m(f_m)) \cos(2\pi f_m (t - \theta)) \tag{3.2}$$

where I_o and I_i are respectively the output of speech intensities and the input of speech intensities, θ is the phase of modulation signal, f_m is the frequency of modulation signal, and $m(f_m)$ is the modulation index of the temporal power envelope that also called MTF.

3.1.2 Speech signal modeling based on MTF

For the speech signal, if $y(t)$, $n(t)$, $h(t)$, and $x(t)$ respectively represent the output speech, noise, room impulse response (RIR) and input speech in an

acoustic room. For the MTF concept, the relation of $y(t)$, $n(t)$, $h(t)$, and $x(t)$ can be represented as

$$y(t) = n(t) + x(t) * h(t) \tag{3.3}$$

$$n(t) = c_n(t)e_n(t) \tag{3.4}$$

$$h(t) = e_h(t)c_h(t) = ac_h(t)\exp(-6.9t/T_R) \tag{3.5}$$

$$x(t) = c_x(t)e_x(t) \tag{3.6}$$

where $e_x(t)$, $e_h(t)$, and $e_n(t)$ respectively represent the temporal envelope of $x(t)$, $h(t)$, and $n(t)$. $c_x(t)$, $c_h(t)$, and $c_n(t)$ respectively represent the carriers signal of $x(t)$, $h(t)$, and $n(t)$ (gaussian white noise). Lastly, T_R represent reverberation time. Based on the theory of stochastic analysis, the following formula is derived

$$\left\langle y^2(t)\right\rangle = \left\langle h^2(t) * x^2(t)\right\rangle + \left\langle n^2(t)\right\rangle \tag{3.7}$$

In this formula,$\langle c_l(t), c_l(t-\tau)\rangle = \delta(\tau)$ with $c_l(t) \in \{c_x, c_h, c_n\}$, $*$ is the convolution calculation, and $\langle\cdot\rangle$ represents an ensemble average operation.

3.1.3 The MTF of complex conditions

The complex MTF can be simulated under noisy, reverberant and noisy reverberant environments. If there is just only reverberate, the complex MTF is defined as the follow function

$$m_R(f_m) = \left[1 + \left(2\pi f_m \frac{T_R}{13.8}\right)^2\right]^{-1/2} \tag{3.8}$$

where f_m is the modulation frequency. The value of MTF is affected by T_R. The larger the T_R, the smaller the MTF.

Like the reverberant environments, the complex MTF in the noisy conditions is defined as

$$m_N(f_m) = \frac{\overline{e_x}^2}{\overline{e_x}^2 + \overline{e_n}^2} = \frac{1}{1 + 10^{-\frac{\mathrm{SNR}}{10}}} \tag{3.9}$$

The value of MTF is affected by SNR. The larger the SNR, the larger the MTF.

If we consider the effects of additive noise and RIR, MTF in the noisy reverberant environment can be calculated as the following function

$$m(f_m) = m_R(f_m) \cdot m_N(f_m) = \frac{1}{\sqrt{1 + \left(2\pi f_m \frac{T_R}{13.8}\right)}\left(1 + 10^{-\frac{\mathrm{SNR}}{10}}\right)} \tag{3.10}$$

As mentioned above, MTF is affected by three parameters: T_R, SNR, and f_m. This means that the low-pass characteristic is produced by the RIR as a function of T_R, and the constant attenuation is produced by the additive noise. In theory, if we know the final MTF, the effect of additive noise and RIR on the noisy reverberant speech temporal envelope can be eliminated by an inverse filter.

3.1.4 Power Envelope Restoration from Complex Conditions

As mentioned in the previous section, if we know the MTF, we can get the clean speech power envelope signal through an inverse filter. In this part, we will introduce in detail the process of using the inverse filter to restore the noisy reverberant speech temporal power envelope. Figure 3.1 is a detailed flowchart of this process, it contains: (i) temporal power envelope feature extraction (ii) temporal power envelope restoration under noisy conditions (iii) temporal power envelope restoration under reverberant conditions.

The temporal power envelope feature is calculated by the following formula

$$e_y^2 = \mathbf{LPF}\left[|y(t) + j\mathbf{Hilbert}(\mathbf{y}(\mathbf{t}))|^2\right] \tag{3.11}$$

where $\mathbf{LPF}\langle\cdot\rangle$ is low-pass filter which the cut-off frequency is 20Hz and $\mathbf{Hilbert}\langle\cdot\rangle$ is Hilbert transform operation.

As for the temporal power envelope restoration under noisy conditions stage to suppress additive noise effect. This step is calculated by the following function

$$\hat{e}_x^2(t) = \bar{e_y^2}\left(1 + m_N(f_m)\cos\left(2\pi f_m t \cdot \frac{1}{m_N(f_m)}\right)\right) = e_y^2 - \bar{e_n^2} \tag{3.12}$$

where $\bar{e_y^2}$ and $\bar{e_n^2}$ respectively represent the mean power of e_y^2 and e_n^2. $\hat{e}_x^2(t)$ is the denoised temporal power envelope. The key to this step is to find the SNR. It is very difficult to get an accurate SNR. How to calculate the accurate SNR in the next section will be explained in detail.

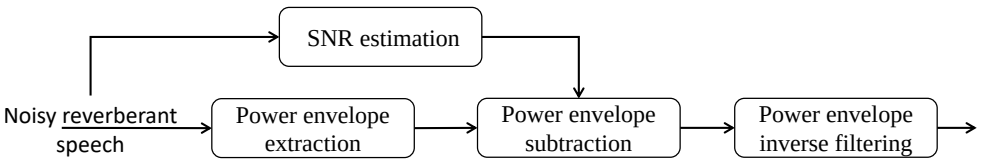

Figure 3.1: Time power envelope recovery of complex conditions speech

For the temporal power envelope restoration under reverberant conditions stage to suppress reverberate effect. This step is calculated from

$$E_x(z) = \frac{E_y}{z}\left\{1 - \exp\left(-\frac{13.8}{T_R \cdot f_s}\right) z^{-1}\right\} \tag{3.13}$$

where $E_x(z)$ represent the z-transforms of dereverberated temporal power envelope, $E_y(z)$ represent the z-transforms of e_y^2, and f_s is the sample frequency. a and T_R are estimated as following function

$$\hat{T}_R = \arg\min\left\{\frac{\mathrm{d}T_P(T_R)}{\mathrm{d}T_R}\right\} \tag{3.14}$$

$$T_P(T_R) = \min\left(\arg\min\left|\theta - \hat{e}_{x,n,T_R}(t)^2\right|\right) \tag{3.15}$$

$$\hat{a} = \sqrt{1/\int_0^T \exp(-13.8t/\hat{T}_R)dt} \tag{3.16}$$

In this research, we assume that the environment is noisy. So, we just only to do denoise for the noisy speech temporal power envelope can restore the original clean speech temporal power envelope.

3.2 Sub-band based gSNR estimation

To eliminate the effect of noise signal, we should calculate the accurate global signal-to-noise ratio (gSNR). Because the additive noise and clean speech components are mixed together, estimating the gSNR in the original time domain signal is very difficult. In previous research, Morita et al. proposed a robust gSNR estimation method [39], which mainly includes sub-band speech signal processing, VAD, and threshold optimization.

The sub-band-based processing method makes speech and noise processing more accurate than global full-band processing. In addition, in previous studies, a fixed threshold is often used to do a final decision of the speech or non-speech parts, but this judgment is not reasonable because different sentences may be in different noise environments. In this study, an optimal threshold is designed to detect speech and noise (for example, different signal-to-noise ratios) under all test conditions to solve the above problems. The optimal threshold is based on minimizing the root mean square (RMS) of the false acceptance rate (FAR) and false rejection rate (FRR) in each sub-band. Finally, gSNR is obtained by calculating the speech energy and noise energy of each sub-band and then calculating the energy ratio of speech and noise. The detailed block of this method will be described which contains sub-band filter design and how to estimate the threshold. Figure 3.2 is the flowchart of this method.

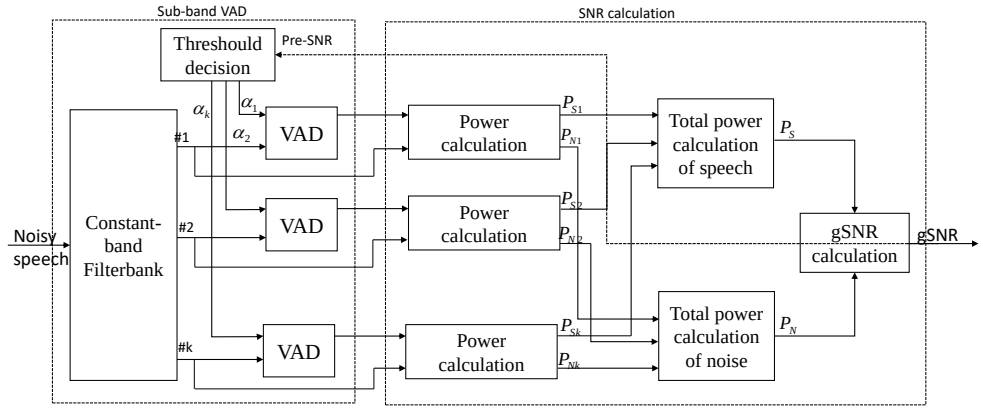

Figure 3.2: Diagram of signal flow with sub-band.

3.2.1 gSNR definition

gSNR refers to the ratio of speech to noise in a speech signal. The speech signal here refers to a signal containing only the additive noise and clean speech refers to an irregular extra signal that does not exist in the original target speech signal generated after a long distance or interference from a recording device. We usually calculate gSNR with the following formula

$$\text{gSNR} = 10 \log_{10} \left(\frac{P_S}{P_N} \right) \tag{3.17}$$

where P_S and P_N respectively represent the target speech waveform and interference noise waveform. Because the energy of speech and noise is not known from the observed noisy speech waveform, they must be estimated based on the detected speech and non-speech period.

3.2.2 Filter-bank design

It is not easy to distinguish noise from speech using the global full-band method. To overcome this, in this study we use a multi-sub-band approach. Noise has different distributions at different frequencies. At high gSNR, because the noise is mostly distributed at high frequencies, the high-frequency components of the sub-bands can be used to easily distinguish the noise. At low gSNR, it is difficult to calculate the energy of speech and noise from noisy speech. The use of sub-bands can process the noise into frequency bands, which makes it easier to judge speech and non-speech parts. Separating the noisy speech into sub-bands of different frequencies can improve the discrimination between noise and speech.

As shown in Figure 3.2., in this study, constant-bandwidth based filter-bank CBFB was used to split the original speech signal into different sub-bands. The CBFB filter-bank consists of a band-pass filter with constant bandwidth. In this research, the bandwidth frequency is set as 100Hz. Since the speech sampling rate is 8,000 Hz, the number of bandwidth filters is 40. Finally, the original noisy speech is split into 40 sub-bands. Lately, by comparing the energy of the sub-band and the estimated threshold, the speech and non-speech parts will be detected from each sub-band separately. The resulting sub-band VAD will be used to calculate its clean speech energy and noise energy to obtain the gSNR.

3.2.3 Threshold decision

After sub-band processing, we get the sub-band speech signal. Since the proportion of speech and noise energy is different in every sub-band, this research detect speech and noise periods in each sub-band use a given power level threshold for each of them. Different sub-bands and different utterances have different thresholds. These decision thresholds are designed based on minimizing the RMS values of FAR and FRR on the receiver operating characteristic (ROC) curve. The local speech energy and noise energy could be accurately estimated by the sub-band FAR and FRR.

The FAR(α) and FRR(α) of different sub-bands are determined by the VAD of the selected threshold in this sub-band. Use these FAR(α) and FRR(α) pairs could further get a ROC curve. In order to more clearly represent these sub-band parameters, these sub-band thresholds could be rewritten as k and the sub-band FAR(α) and FRR(α) could be rewritten as FAR(α_k) and FRR(α_k). This study uses a noisy data corpus to train under different gSNR conditions and noise types to find the most optimal FAR(α_k) and FRR(α_k) pairs. The objective function of finding the optimal sub-band threshold is determined by minimizing the RMS(α_k) obtained from FAR(α_k) and FRR(α_k) which could be written as

$$\alpha_k^* = \arg\min \mathrm{RMS}(\alpha_k) \tag{3.18}$$

$$\mathrm{RMS}(\alpha_k) = \sqrt{\frac{\mathrm{FAR}^2(\alpha_k) + \mathrm{FRR}^2(\alpha_k)}{2}} \tag{3.19}$$

After obtaining the optimal sub-band thresholds at all gSNRs, the sub-band threshold-gSNR curve is fitted, As shown in Figure 3.3. To better fit the sub-band gSNR-threshold curve, this study used a fourth-order Sigmoid function. In the fitting function, the minimum mean square error of the best threshold value and the threshold value obtained under the true gSNR are used as the fitting criterion.

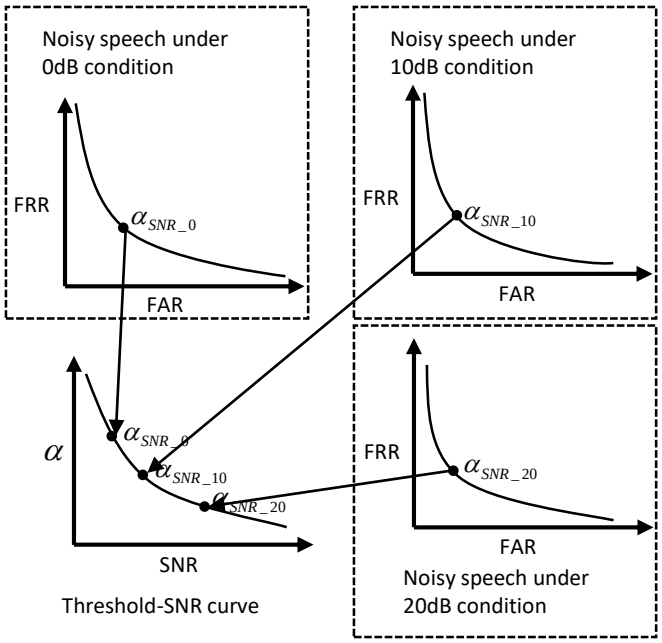

Figure 3.3: The estimation of threshold-SNR curve

3.2.4 Power calculation use for sub-band threshold

The corresponding sub-band VAD can be obtained through the sub-band threshold. The final gSNR is calculated by the following formula

$$\mathrm{gS\hat{N}R} = 10\log_{10}\left(\frac{\sum_{k=1}^{K} P_{STk}}{\sum_{k=1}^{K} P_{NTk}}\right) \tag{3.20}$$

$$P_{NTk} = \int_0^T \overline{P_{Nk}} H_{Sk}(t)dt \tag{3.21}$$

$$P_{STk} = \int_0^T P_{SNk} H_{Sk}(t)dt - \int_0^T \overline{P_{Nk}} H_{Sk}(t)dt \tag{3.22}$$

where P_{NTk} and P_{STk} are the energy of the sub-band noise and speech, respectively. K is the sub-bands number, Hsk is the VAD decision under different sub-band.

In addition to the above process, this study also used an iterative approach to adjust the estimated gSNR. In this loop, the estimated SNR is fed to the threshold decision phase of the VAD. Then, reset the decision threshold of VAD for the gSNR estimate in the next iterations. After multiple iteration, the estimated gSNR is converged to the convergence point of the gSNR of the loop based on the threshold-gSNR curve.

Chapter 4

Proposed Method

The utility of the MTF concept for the VAD task has been proposed to improve robustness against noisy conditions [20]. In this method, the concept of MTF has been used to reduce the bad affect of additive noise on the target speech. However, the robustness of the method decreases under realistic conditions in which non-stationary and low SNR noise conditions appear. Besides, methods based on deep neural networks (DNNs) have been proposed to directly learn nonlinear functions to do VAD as an end-to-end model [35]. However, the DNNs-based method must have massive datasets including these noises. If the speech collected in the real life is not included in these environments, the model needs to be retrained. This research aims to solve the above two issues by incorporating the MTF concept into DNNs architecture. Specifically, in order to solve the MTF based method decreases under realistic conditions, we make efforts to the exploration of improving gSNR accurate which use for the CNN encoder-decoder. Then the estimated gSNR is used to do a restoration of the temporal power envelope. This chapter will display the overall framework of the proposed method and then explain the gSNR estimation method based on CNN encoder-decoder in detail.

4.1 Framework for proposed VAD method

Similar to Morita's method, This research also propose a temporal power envelope based VAD method. To achieve robust VAD, we must remove the bad affect of noise on the power envelope. In this study, we propose a method for restoring envelope features use for the MTF to release the effect of additive noise under noisy environments for speech processing.

As shown in Figure 4.1, the proposed framework have the envelope fea-

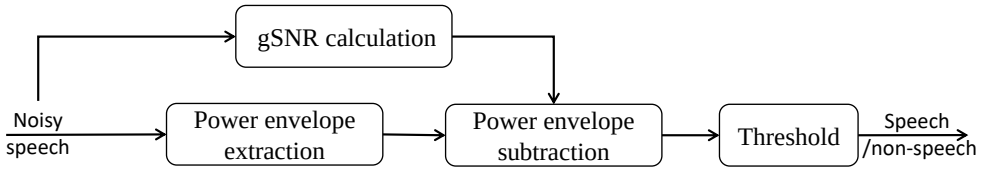

Figure 4.1: Framework for proposed VAD method.

ture extraction, gSNR estimation, the temporal power envelope restoration and the speech/non-speech decision step. In the previous work, a similar framework has already used for achieving a robust VAD performance under reverberant environments. to remove the bad affect of additive noise, we have done a envelope subtraction calculation. The key gSNR will be estimated to do power envelope subtraction. Finally, by using a threshold the speech/non-speech will be decided. In the following section, we will give a detail introduction to the proposed method.

4.2 Restoration of the temporal power envelope

For the proposed method, the first step is extract temporal power envelope feature. The power envelope feature extracted from the following function

$$e_y^2(t) = \mathbf{LPF}\left[|y(t) + j\mathbf{Hilbert}(y(t))|^2\right] \tag{4.1}$$

where $\mathbf{LPF}[\cdot]$ represent the low pass filter of cut-off frequency is 20Hz and $\mathbf{Hilbert}[\cdot]$ represent Hilbert transform.

If there is just only additive noise, from the last chapter the MTF could be represented as

$$m_N = \frac{1}{1 + 10^{-\frac{gSNR}{10}}} \tag{4.2}$$

$$e_y^2(t) = e_x^2(t) + \overline{e_y^2(t)} \cdot (1 - m_N) \tag{4.3}$$

where e_x^2 represent the clean speech power envelope. How to estimate the gSNR will be introduced in the following paper. We can design an inverse filter corresponding to m_N to remove the bad affect of additive noise on the noisy speech temporal envelope feature. The restoration of the temporal power envelope can be calculated as

$$\hat{e}_x^2(t) = e_y^2(t) - \overline{e_y^2(t)} \cdot (1 - m_N) \tag{4.4}$$

where $\hat{e}_x^2(t)$ is the envelope feature by remove the additive noise bad affect.

4.3 Sub-band based dynamic SNR estimation method

To eliminate the bad affect of additive noise, we should estimate the gSNR in the Equation 4.3. As shown in the last section, Morita et al. proposed a robust gSNR detection method in the sub-band speech signal [39]. But with the increase of noise (decreased SNR), because VAD cannot usually be accurately judged by a single threshold in the whole sentence, this method is not robust at low SNR. Many people have proposed DNNs based end-to-end gSNR detection methods [41, 42], end-to-end based methods usually extract acoustic features of one utterance and then input all of them to a neural network to predict gSNR. This method often has the problem of data mismatch or environment mismatch. In addition, the end-to-end method requires that the input utterance have the same shape. If they are different, you need to do pooling and speech cutting to make all sentences the same length. This kind of processing method will cause the problem of out of memory and cannot adapt to all applications.

To solve the above problems, we propose an indirect gSNR estimation method. Because all additive noise and clean speech components are mixed in one noisy speech, it is not easy to calculate the gSNR in the original speech waveform. In this research, we use for the sub-band signal processing method. The proposed gSNR estimation method mainly includes sub-band speech signal processing, sub-band threshold calculation unit, sub-band power calculation unit and gSNR calculation unit. The sub-band-based processing method makes speech and noise processing more accurate than global full-band processing. In addition, in previous studies, a static threshold often does the final VAD decision, but this judgment is not reasonable because different speech samples may be in different threshold or different noise ratios. In this study, based on the sub-band and CNN encoder-decoder structure we propose a gSNR estimation method, this method could estimate the noise ratio of different sub-band speech signal sample. Figure 4.2 is a diagram of the proposed method.

4.3.1 Sub-band processing design

It is not easy to distinguish noise from noisy speech using the global full-band method. To overcome this, in this study we use a multi-sub-band approach. Noise has different distributions at different frequencies. At high gSNR conditions, because the noise is mostly distributed at high frequencies, the high-frequency components of the sub-bands can be used to easily distin-

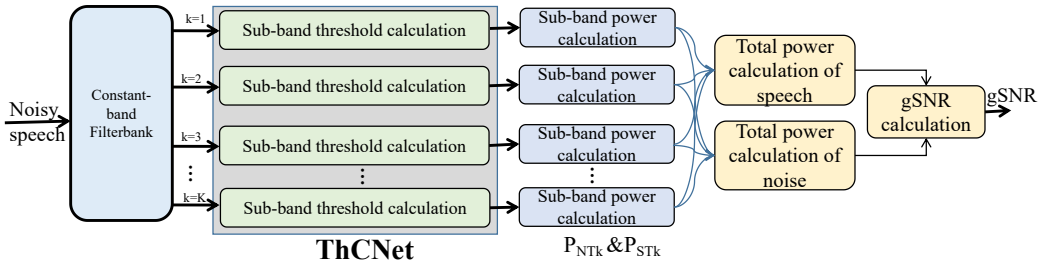

Figure 4.2: Block diagram of signal flow with proposed gSNR method.

guish the noise. At low gSNR conditions, it is difficult to estimate the energy of speech and noise from noisy speech. The use of sub-bands can process the noise into frequency bands, which makes it easier to determine speech and noise parts. Separating the noisy speech into sub-bands of different frequencies can improve the discrimination between noise and speech.

As shown in Figure 4.2, in this study, constant-bandwidth based filter-bank CBFB was used to split the original speech waveform into different sub-bands. The CBFB filter-bank consists of a band-pass filter with constant bandwidth. In this research, the bandwidth frequency is set as 200Hz. Since the speech downsamples to 8,000 Hz, the number of bandwidth filters is 20. Finally, the original speech is split into 20 sub-bands. This process could be represented in the following function

$$[y_1(t), y_2(t), ..., y_n(t)] = \text{CBFB}(y(t)) \tag{4.5}$$

where CBFB[$\cdot$] is the constant bandwidth filter-bank and $[y_1(t), y_2(t), ..., y_n(t)]$ is the sub-band speech signal.

Lately, by comparing the energy of the sub-band and the estimated threshold, the speech and non-speech parts will be detected from each sub-band separately. The sub-band speech signal energy could be computed as the following function

$$[E_1(t), E_2(t), ..., E_n(t)] = [|y_1(t)|^2, |y_2(t)|^2, ..., |y_n(t)|^2] \tag{4.6}$$

where $E_1(t), E_2(t), ..., E_n(t)$ is the sub-band speech energy. The resulting sub-band VAD will be used to calculate its clean speech energy and noise energy to obtain the gSNR.

4.3.2 Threshold calculation network

After got the sub-band speech signal is the sub-band threshold calculation stage. Figure 4.3 is the diagram of the proposed sub-band threshold calculation network. First is the threshold calculation network (ThCNet) training

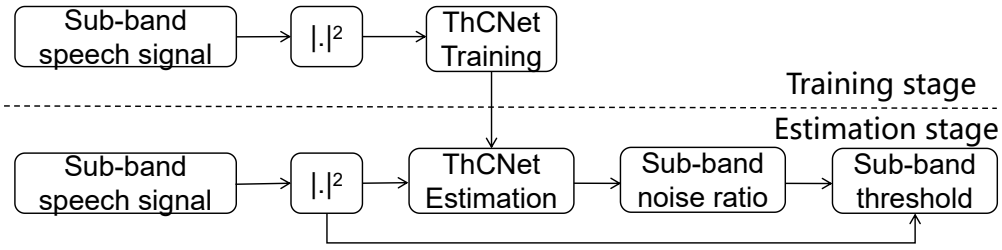

Figure 4.3: The diagram of proposed sub-band threshold calculation.

stage, the ThCNet will be trained from the noisy speech sub-band waveform energy and sub-band noise ratio. Then the well trained ThCNet is fed with the features of noisy speech sub-band speech signal energy for the generation of the sub-band noise ratio. Afterward, the sub-band noisy speech energy is multiplied by the corresponding sub-band noise ratio to get the sub-band threshold.

The context of the speech signal is related, how to correlate the correlation between these contexts can improve the accuracy of the model. To exploit more accurate context information from the given sub-band speech signal energy, CNN is used in the ThCNet. ThCNet can fully explore the characteristics of the sub-band signal, which will greatly promote the accuracy of the sub-band noise ratio, then get an accurate sub-band threshold. Unlike Morita's method, this method makes full use of the advantages of nonlinear learning of neural networks to learn a dynamic threshold learning method. It will let our proposed method more robust under low SNR conditions. In addition, since the model based on deep learning is trained under many different kinds of noise, the proposed method can adapt to many different kinds of noise.

Formally, $\mathbf{E} = E_1(t), E_2(t), ..., E_n(t)$ represents the noisy speech sub-band energy and $\mathbf{R} = r_1, r_2, ..., r_n$ represents its corresponding sub-band noise ratio which is calculated by the following function

$$[r_1, r_2, ..., r_n] = \left[\frac{\int_0^\infty EN_1(t)dt}{\int_0^\infty E_1(t)dt}, \frac{\int_0^\infty EN_2(t)dt}{\int_0^\infty E_2(t)dt}, ..., \frac{\int_0^\infty EN_n(t)dt}{\int_0^\infty E_n(t)dt} \right] \tag{4.7}$$

$$[EN_1(t), EN_2(t), ..., EN_n(t)] = \left[|n_1(t)|^2, |n_2(t)|^2, ..., |n_n(t)|^2 \right] \tag{4.8}$$

where $n_n(t)$ is the sub-band waveform of noise, $EN_n(t)$ is the the sub-band noise waveform energy. Given a training set of sub-band noisy and sub-band noise ratio, the problem of sub-band noise ratio estimation is formalized as finding a mapping that maps a noisy speech energy to a noise ratio $g_\theta(\mathbf{E})$. Then the following optimization problem is solved for obtaining the best

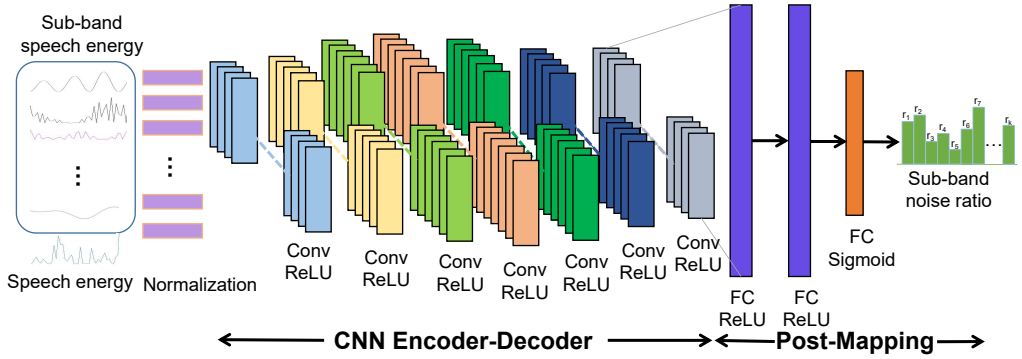

Figure 4.4: The proposed CNN encoder-decoder based sub-band noise ratio estimation structure

model parameter θ

$$\hat{\theta} = \arg\min \sum_{i=1}^{n} \|g_\theta(\mathbf{E}) - \mathbf{R}\|^2 \tag{4.9}$$

where n is the sub-band number. Under this setting, ThCNet is designed as mapping function $g_\theta(\mathbf{E})$ for noise ratio estimation. As shown in the Figure 4.4, it is mainly making up by CNN encoder-decoder, post-mapping and the output composition.

CNN encoder-decoder component To exploit a more accurate local context pattern from given sub-band speech energy, the CNN encoder-decoder is utilized in ThCNet. Not just only use for the fully connected layer, we use another convolutional network structure which name is the CNN encoder-decoder (C-ED) network. As shown in Figure 4.4, C-ED is made up of convolution, batch normalization, and ReLU layer. Because the pooling layer always leads to the loss of information, there is no pooling and upsampling layer in the C-ED. The number of encoder and decoder filters is corresponding, the number of encoder filters gradually increases, and the number of decoder filters gradually decreases. To do the generalization ability improvement of the model, we set different convolution kernels in the CNN model to learn different context pattern. By the C-ED, the hidden representation $V(\mathbf{E})$ of target sub-band noise ratio is generated as

$$V(\mathbf{E}) = \mathrm{CED}(\mathbf{E}) \tag{4.10}$$

Post-mapping component In order to estimate noise more accurately, a fully connected layer-based network is used in the ThCNet. Through deeper non-linear operations, the network can predict more detailed information,

which is conducive to the learning of the sub-band noise ratio. The post-mapping network consists of two layers full connected layers in which the activation function is ReLU. By the post mapping network, the hidden representation of target sub-band noise ratio is generated as

$$M(\mathbf{E}) = \text{ReLU}(\mathbf{W_2}(\text{ReLU}(\mathbf{W_1}V(\mathbf{E}) + \mathbf{b_1})) + \mathbf{b_2}) \tag{4.11}$$

$$\text{ReLU}(x) = \max(0, x) \tag{4.12}$$

Output composition In this layer by a Sigmoid function will let the output noise ratio $[\hat{r}_1, \hat{r}_2, ..., \hat{r}_n]$ to the around 0 to 1

$$[\hat{r}_1, \hat{r}_2, ..., \hat{r}_n] = \text{Sigmoid}(\mathbf{W}_3 M(\mathbf{E}) + \mathbf{b}_3) \tag{4.13}$$

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{4.14}$$

In the threshold estimation stage, we use for the following equation to calculate the threshold

$$[T_1, T_2, ..., T_n] = \left[\hat{r}_1 \int_0^\infty E_1(t)dt, \hat{r}_2 \int_0^\infty E_2(t)dt, ..., \hat{r}_n \int_0^\infty E_n(t)dt\right] \tag{4.15}$$

4.3.3 gSNR calculation

The power of a speech waveform in the time domain is calculated from the summation of power from all sub-bands. Since threshold has been separately designed in each sub-band, the estimates of noise and speech powers are much more accurate than direct estimates in the time domain. The final gSNR is obtained from the power fusion of all sub-bands as

$$\hat{\text{gSNR}} = 10 \log_{10} \left(\frac{\sum_{k=1}^{N} P_{STk}}{\sum_{k=1}^{N} P_{NTk}} \right) \tag{4.16}$$

$$P_{NTk} = \frac{\sum_{i=1}^{LN} T_k^i}{LN} \cdot L(\hat{r}_k > P) \tag{4.17}$$

$$P_{STk} = \sum_{i=1}^{LN} E_k^i - \frac{\sum_{i=1}^{LN} T_k^i}{LN} \cdot L \tag{4.18}$$

where P_{NTk} and P_{STk} are the total power of additive noise and clean speech in the k-th sub-band, LN is the number of thresholds T_k^i when the sub-band noise ratio bigger than the parameter P. L is the total length of utterance.

Using a C-ED structure can fully learn the relationship between contexts in speech features, making acoustic feature learning more fully. In addition,

since the model at the same time inputs different types of noisy speech into the network for training, the proposed method can adapt to thresholds under a variety of different types of noise. Compared with Morita's method, it is not necessary to adjust the model according to the type of noise. In addition, since the proposed method can dynamically learn the threshold in each utterance. Because the Morita's static threshold method based on simple ROC curve only seeks the average value of VAD accuracy suitable for the whole utterance, it cannot fully use the information of the acoustic feature itself. Therefore, compared with the method proposed in the previous method, better gSNR performance can be obtained. Substituting this better gSNR into Equation 4.3 can better restore the temporal power envelope of speech, and further get better VAD performance.

Chapter 5

Evaluation

5.1 Dataset

To do the final evaluation, we use for the speech data of AURORA-2J [43] and NOISEX-92 [44] dataset. Inthe training dataset, 8440 clean speech utterances of AURORA-2J were selected as clean. White, pink, factory and babble noise in NOISEX-92 were used as background noise like Table 5.1. Noisy speech signals were artificially created as

$$y(t) = n(t) + x(t) \tag{5.1}$$

where $y(t)$ is noisy speech, $n(t)$ is the background noise, and $x(t)$ is the clean speech signal. Noisy speech signals with SNRs of 20, 15, 10, 5, 0, -5, and -10 were generated like the Table 5.2. These clean and noisy speech signals were then used to find the SNR design. The sampling frequency was 8 kHz, the bandwidth of sub-bands was 200 Hz, and the number of sub-bands was 20.

Table 5.1: Noise type used in the experiments

Noise type			
White	Pink	Factory	Babble

Table 5.2: SNR type used in the experiments

SNR						
20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB	−10 dB

5.2 Experimental setup

Our gSNR calculation is finally used to do the noisy speech temporal power envelope denoise. Tensorflow is used to train our ThCNet with a CNN encoder-decoder model and the sub-band energy features. All of the hidden layer use for the *ReLU* as the activation function. We used the *Adam* algorithm [45] as optimizer. The convolution layer filters number is 21, 40, 64, 128, 64, 40, 21 and the kernel size is set as 2, 3, 5, 7, 5, 3, 2. In the Mapping-net, the hidden size is set as 512. Lastly, the batch size is set to 64, the learning rate of our method is set to 0.01.

The evaluation criterion of gSNR is the mean absolute error (MAE) for the estimated gSNR and the real gSNR, the mean estimated gSNR of every utterance also is used to do gSNR estimation.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|G_i - R_i| \tag{5.2}$$

$$\text{MEAN} = \frac{1}{N}\sum_{i=1}^{N}G_i \tag{5.3}$$

where N is the number of test dataset (1001), G is calculated gSNR, and R is real gSNR. The MAE lower, the better gSNR performed. To evaluate the performance of VAD, the RMS (%) of FRR (%) and FAR (%) is used as

$$\text{RMS} = \sqrt{\frac{\text{FRR}^2 + \text{FAR}^2}{2}} \tag{5.4}$$

and a smaller RMS indicates better results.

5.3 gSNR results and analysis

Figure 5.1 is the average value of estimated gSNR and previous gSNR estimation methods under different noise environments and different SNR. The analysis results display that the C-ED based method is closer to the ideal gSNR than the method proposed by Morita's gSNR method. The table reveal that the estimated gSNR of Morita's method is generally higher than the ideal value in a low SNR environment. This is because the gSNR method of Morita uses a threshold to determine VAD, but it is not reasonable to use only the threshold to judge VAD, because the speech and non-speech waveform are very close in a low SNR environment. If just only use a fixed threshold to judge speech and non-speech for an utterance, the VAD results

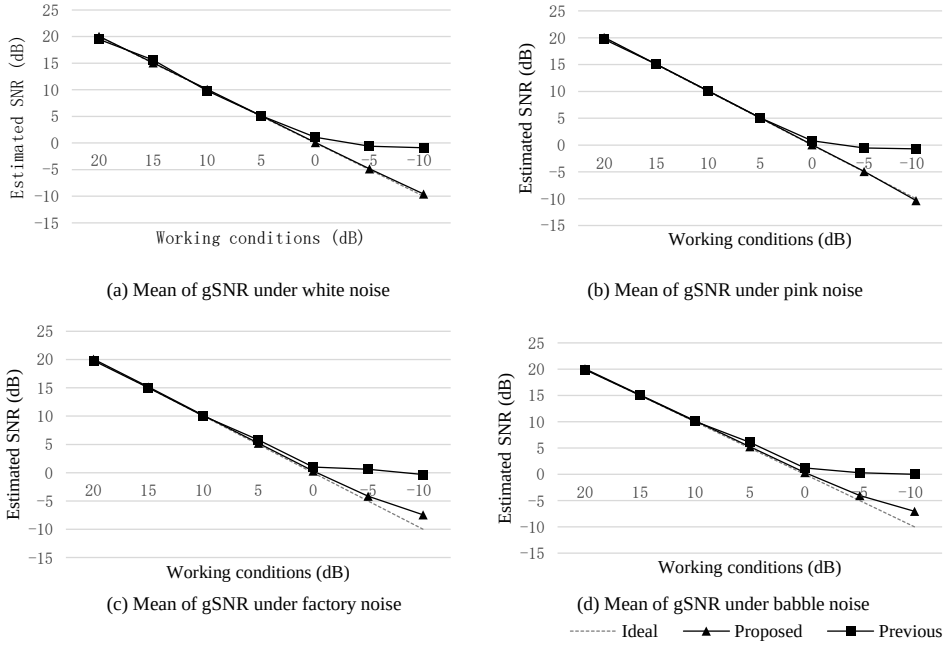

Figure 5.1: Mean of estimated gSNR under different noise conditions

judged at low SNR will tend to be almost random. Our proposed method uses deep learning to learn a dynamic noise ratio in a short speech frame. Deep learning can learn this nonlinear relationship well. This makes the learned threshold dynamic rather than static in an utterance. Therefore, the proposed method has a significant performance improvement over Morita's method, it is nearly close to ideal values especially with white and pink noise conditions. For some non-stationary noise, the proposed method can achieve the ideal value when the SNR is greater than 0. When the SNR is less than 0, although the result is greatly improved compared with the previous method, the non-stationary noise is irregular due to the change in utterance, which is very similar to real speech. The estimation of gSNR under non-stationary noise is still a difficult problem.

In order to detect the proposed method more fairly, as shown in Figure 5.2, we also use MAE as the detection index. Using MAE can more objectively evaluate the performance of a single sample. The mean is more representative of the performance of the overall sample. From the MAE analysis, it reveal that the C-ED based method has a smaller value than the Morita's method. This shows that the performance of the proposed method is more stable than that of Morita's method, which will enable the proposed method to adapt to more and more complex environments. The proposed

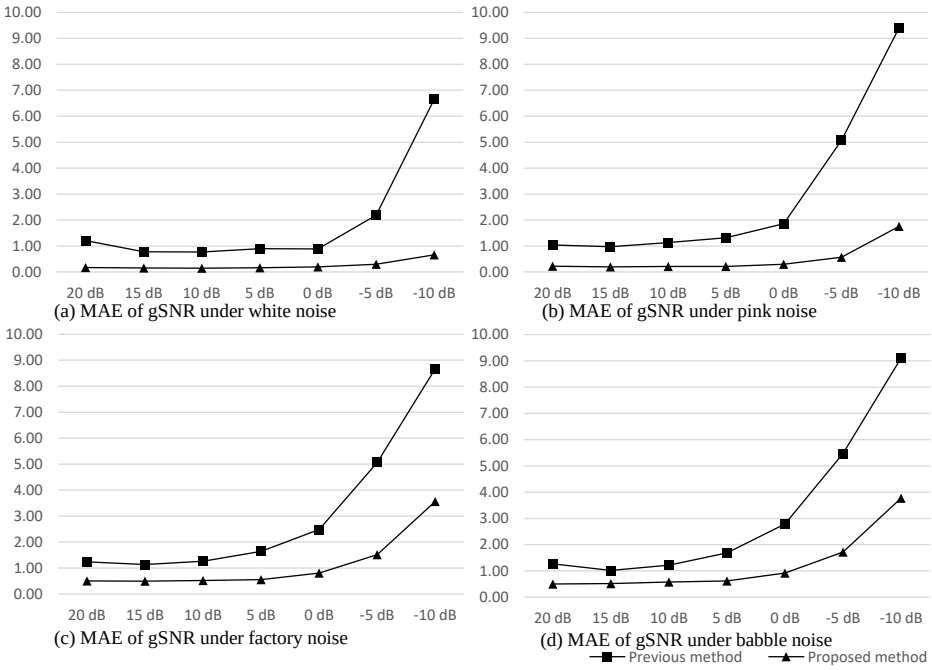

Figure 5.2: MAE of gSNR under different noise conditions

method has stable performance under white noise and pink noise, but under non-stationary noise, although it has been greatly improved compared to the Morita's method, its MAE size is still large. This shows that the results under unstable noise conditions are unstable. This is because the proposed C-ED based method calculates the average value of the energy of a speech segment below a certain threshold and then uses the total energy and its subtraction to obtain the energy of its speech segment. The energy distribution of non-stationary noise in a speech is unstable, this will make the result not ideal. The use of sub-band processing can alleviate this problem to some extent, but how to improve performance under non-steady-state noise is still quite difficult.

5.4 VAD results and analysis

Figure 5.3 shows the VAD results under white noise. The results revealed that the previous method has a performance degradation in the low SNR environments. Although the performance of the proposed method decreases in a low SNR environment, the VAD results of the proposed C-ED based VAD method are significantly improved compared to previous methods due

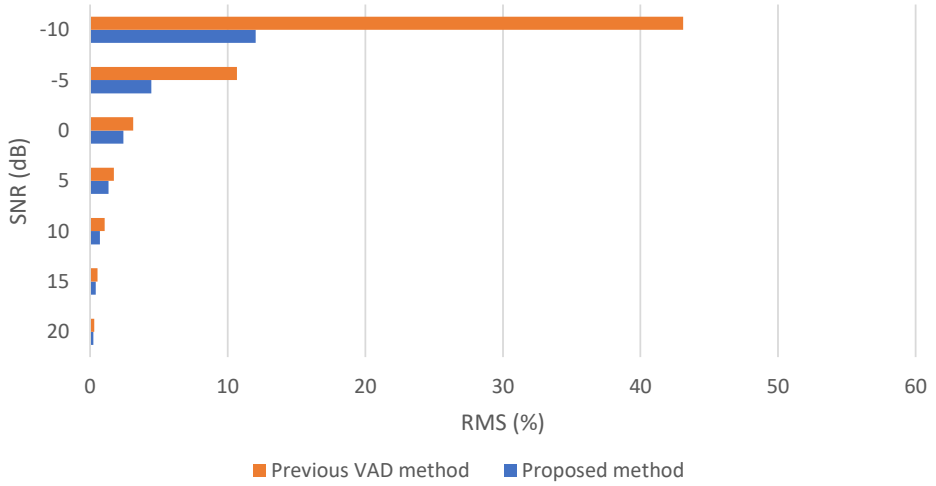

Figure 5.3: VAD Results for accuracy of RMS(%) under white noise

to the very accurate gSNR estimation of the C-ED based VAD proposed method. The accuracy of the proposed C-ED based VAD method is still maintained at about 95% RMS at a very low SNR.

Figure 5.4 shows the results of VAD under pink noise. The results reveal that the performance of the previous method or the proposed method is reduced under pink noise. This is because pink noise has some unstable phonemes. In addition, the previous method has significant performance degradation in a low SNR environment. Due to the very accurate gSNR estimation of the proposed C-ED based method, the VAD results are significantly improved compared to previous methods. The proposed C-ED based VAD method is still maintained at about 90% RMS at very low SNR.

Figure 5.5 shows the results of VAD under factory noise. The results reveal that the performance of both the previous method and the proposed method is reduced under the factory noise. This is because factory noise is an unstable noise. As in the case of white noise and pink noise, the previous method has a significant performance degradation in low SNR environments. Due to the very accurate gSNR estimation of the proposed method, the VAD results are significantly improved compared to previous methods. Since the factory is non-stationary noise, its energy characteristics are very similar to speech, so VAD detection under this condition is very difficult. The difference between white and pink in this environment is that the latter has a more average energy distribution, which makes the estimation of gSNR more difficult than the unstable factory noise. This is an important reason why the performance of both the previous method and the proposed method is

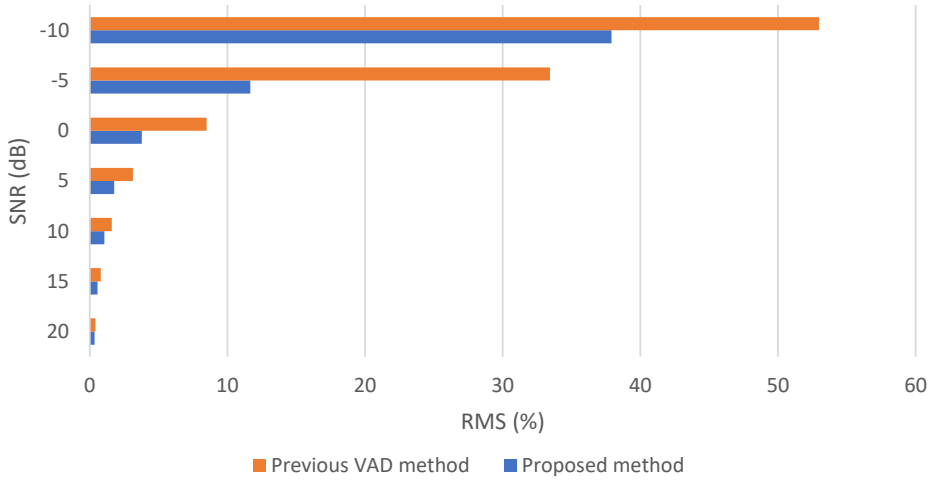

Figure 5.4: VAD Results for accuracy of RMS(%) under pink noise

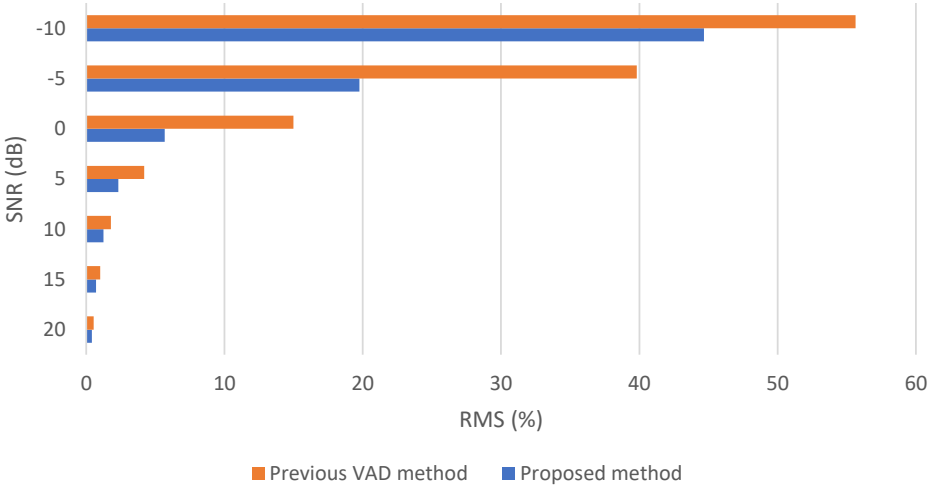

Figure 5.5: VAD Results for accuracy of RMS(%) under factory noise

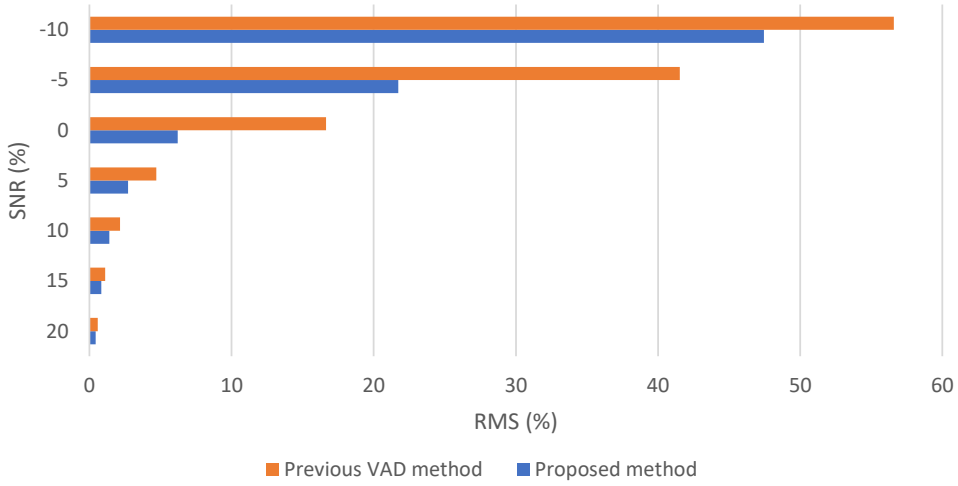

Figure 5.6: VAD Results for accuracy of RMS(%) under babble noise

rapidly degraded.

Figure 5.6 shows the results of VAD under babble noise. This result is similar to that in the factory environment. This is because babble noise is also unstable noise. How to solve the VAD problem in low SNR and unstable noise environment is still very difficult.

Chapter 6

Conclusion

6.1 Summary

This research proposed a sub-band based dynamic gSNR estimation method used convolutional neural network encoder-decoder for the power envelope restoration. Further, this restored power envelope was used to do final speech/non-speech decision. Comparing with the previous modulation transfer function based voice activity detection method our method could work better under non-stationary noise and low SNR environments.

The advantage of the proposed method has two-point. Sub-band-based speech signal processing can process speech and noise signals in different frequency bands to obtain a robust gSNR, and if changing the environment does not require readjusting model parameters. Secondly, this method outperformed the conventional sub-band based VAD in terms of accuracy.

In conclusion, a voice activity detection system utilizing the convolutional neural network encoder-decoder model has been proved to achieve better performance compared to the previous modulation transfer function based method.

6.2 Contribution

The voice activity detection system utilizing a convolutional neural network encoder-decoder model was proposed in this research. The proposed CNN encoder-decoder based VAD method can improve the accuracy of speech/non-speech detection similar to the previous method using the modulation transfer function. Although there was still much room for performance improvement, this research put the first step toward the realization of incorporating deep neural networks into modulation transfer function concept. Moreover,

this will further provide key technical support for not only various speech applications but also man-machine speech communications under real environmental conditions.

6.3 Remaining works

- Under a very low SNR babble noise environment, the gSNR estimation is still a problem, it will further affect the voice activity detection performance. Therefore, using what feature and concept could solve the above problem is still a direction in the future.
- The reverberant speech also has a bad effect on the VAD, how to eliminate the reverberant speech effect also another direction in the future.
- Proposed gSNR could predict very well under low gSNR. In future work, I would use it in some other speech signal processing method, e.g. Wiener filter.

Bibliography

[1] G. Hinton, L. Deng, and D. Yu, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82-97, 2012.

[2] D. Snyder, D. Garcia-Romero, and G. Sell, "X-vectors: Robust Dnn Embeddings for Speaker Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329-5333, 2018.

[3] L. Guo, L. Wang, and J. Dang, "Speech Emotion Recognition by Combining Amplitude and Phase Information Using Convolutional Neural Network," *Interspeech 2018*, pp. 1611-1615, 2018.

[4] N. Li, M. Ge, and L. Wang, "A Fast Convolutional Self-attention Based Speech Dereverberation Method for Robust Speech Recognition," *International Conference on Neural Information Processing*, pp. 295-305, 2019.

[5] M. Ge, L. Wang, and N. Li, "Environment-dependent Attention-driven Recurrent Convolutional Neural Network for Robust Speech Enhancement," *Interspeech 2019*, pp. 3153-3157, 2019.

[6] H. Veisi and 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.

[7] S. Tong, N. Chen, Y. Qian, and K. Yu, "Evaluating VAD for Automatic Speech Recognition," *12th International Conference on Signal Processing (ICSP)*, pp. 2308-2314, 2014.

[8] Y. Mamiya, J. Yamagishi, and O. Watts, "Lightly Supervised GMM VAD to use Audiobook for Speech Synthesiser," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7987-7991, 2013.

[9] T. Kinnunen, and P. Rajan, "A Practical, Self-adaptive Voice Activity Detector for Speaker Verification with Noisy Telephone and Microphone Data," *IEEE international conference on acoustics, speech and signal processing*, pp. 7229-7233, 2013.

[10] D. K. Freeman, C.B. Southcott, I. Boyd, and G. Cosier, "A Voice Activity Detector for Pan-European Digital Cellular Mobile Telephone Service," *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 369-372, 1989.

[11] M. H. Moattar, and M. M. Homayounpour, "A Simple But Efficient Real-time Voice Activity Detection Algorithm," *European Signal Processing Conference*, 2009.

[12] J. Stegmann, and G. Schroder, "Robust Voice-activity Detection Based on the Wavelet Transform," *IEEE Workshop on Speech Coding for Telecommunications Proceeding*, 1997.

[13] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings I (Communications Speech and Vision)*, vol. 139, no. 4, pp. 377-380, 1992.

[14] ITU, "Coding of Speech and 8 kbit/s Using Conjugate Structure Algebraic Code -Excited Linear Prediction. Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommend," *International Telecommunication Union*, 1996.

[15] 3GPP Organizational Partners, "Adaptive Multi Rate (AMR) Speech; ANSI-C code for AMR Speech Codec," 1998.

[16] ETSI, "Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi Rate (AMR); Speech Processing Functions; General Description," 1998.

[17] B. S. Liu, Z. M. Lu,and L. R. Shen, "Voice Activity Detection with Low Signal-to-noise Ratio Based on Hilbert-Huang Transform," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 41, no. 3, pp. 844-848, 2011.

[18] Y. Kanai, and M. Unoki, "Robust Voice Activity Detection Using Empirical Mode Decomposition and Modulation Spectrum Analysis," *International Symposium on Chinese Spoken Language Processing*, 2013.

[19] M. Unoki, X. Lu, R. Petrick, S. Morita, M. Akagi, and R. Hoffmann, "Voice Activity Detection in MTF-based Power Envelope Restoration," *In Proceedings Interspeech2011*, pp. 2609-2612, 2011.

[20] S. Morita, M. Unoki, X. Lu, and M. Akagi, "Robust Voice Activity Detection Based on Concept of Modulation Transfer Function in Noisy Reverberant Environments," *Journal of Signal Processing Systems*, pp. 108-112, 2015.

[21] F. Hppner, F. Klawonn, and R. Kruse, "Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition," 1999.

[22] C. D. Manning, and H. Schtze, "Foundations of Statistical Natural Language Processing," *MIT press*, 1999.

[23] P. Resnick, and H. R. Varian, "Recommender Systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56-58, 1997.

[24] F. Jelinek, "Statistical Methods for Speech Recognition," 1997.

[25] J. Sohn and W. Sung, "A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation," *IEEE Int. Conf. Acoust., Speech, Signal Process*, pp. 365-368, 1998.

[26] 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.

[27] Y. Li, R. Zhang, and H. Cui, "Voice Activity Detection Algorithm with Low Signal-to-noise Ratios Based on the Spectrum Entropy," *Journal of Tsinghua University*, 2005.

[28] A. Davis, S. Nordholm, and R. Togneri, "Statistical Voice Activity Detection Using Low-Variance Spectrum Estimation and an Adaptive Threshold," *IEEE Trans. Audio Speech Lang. Process*, vol. 14, pp. 412-424, 2006.

[29] Y. C. Lee, and S. S. Ahn, "Statistical Model-Based VAD Algorithm with Wavelet Transform," *IEICE Trans. Fundam. Electron. Commun. Comput*, pp. 1594-1600, 2006.

[30] J. H. Chang, N. S. Kim, and S. K. Mitra, "Voice Activity Detection Based on Multiple Statistical Models," *IEEE Trans. Signal Process*, vol. 54, no. 6, pp. 1965-1976, 2006.

[31] J. M. Grriz, J. Ramrez, and E. W. Lang, "Hard C-means Clustering for Voice Activity Detection," *Speech communication*, vol. 48, no. 12, pp. 1638-1649, 2006.

[32] D. Ying, Y. Yan, and J. Dang, "Voice Activity Detection Based on an Unsupervised Learning Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8 pp. 2624-2633, 2011.

[33] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain," vol. 65, no. 6, pp. 386, 1958.

[34] S. K. Pal, and S. Mitra, "Multilayer Perceptron, Fuzzy Sets, Classifiaction," 1992.

[35] X. L. Zhang, and J. Wu, "Deep Belief Networks Based Voice Activity Detection," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 4, pp. 697-710, 2013.

[36] X. L. Zhang and D. L. Wang, "Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection," *IEEE/ACM Trans. Audio, Speech*, vol. 24, no. 2, pp. 252-264, 2016.

[37] D. A. Silva, J. A. Stuchi, R. P. V. Violato, and L. G. D. Cuozzo, "Exploring Convolutional Neural Networks For Voice Activity Detection," *Cognitive Technologies*, pp. 37-47, 2017.

[38] J. Kim, J. Kim, S. Lee, J. Park, and M. Hahn. "Vowel Based Voice Activity Detection with LSTM Recurrent Neural Network," *8th Int. Conf. Signal Process. Syst.*, pp. 134-137, 2016.

[39] S. Morita, X. Lu, M. Unoki, and M. Akagi, "Method of Estimating Signal-to-Noise Ratio Based on Optimal Design for Sub-Band Voice Activity Detection," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 8, no. 6, pp. 1446-1459, 2017.

[40] S. Hirobayashi, H. Nomura, and T. Koike, "Speech Waveform Recovery from a Reverberant Speech Signal Using Inverse Filtering of the Power Envelope Transfer Function," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 83, no. 6, pp. 77-85, 2000.

[41] P. Papadopoulos, A. Tsiartas, and S. Narayanan "Long-term SNR estimation of speech signals in known and unknown channel conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2495-2506, 2016.

[42] X. Dong, and D. S. Williamson, "Long-term SNR estimation using noise residuals and a two-stage deep-learning framework," *International Conference on Latent Variable Analysis and Signal Separation*, pp. 351-360, 2018.

[43] S. Nakamura, K. Takeda, and K. Yamamoto, "AURORA-2J: An Evaluation Framework for Japanese Noisy Speech Recognition," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 535-544, 2005.

[44] A. Varga, and 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*, vol. 12, no. 13, pp. 247-251, 1993.

[45] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.