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

1810249 LI, Nan

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.