Sonos hire challenge: Voice Activity Detector

Thai Ba Tuan
Master 2 IASD
Paris Dauphine-PSL University
70016, Paris
ba-tuan.thai@dauphine.eu

February 05th, 2022

Abstract

The task of producing a Voice Activity Detector (VAD) that is robust in the presence of non-stationary background noise has been an active area of research for several decades. Historically, many of the proposed VAD models have been highly heuristic in nature. Recent year, Deep Learning method have been expored and used and its have a very powerful when compare with the traditional method. This report introduce how to build, design the Voice Activity Detector model with Deep Learning and apply it to predict voice in the audio file as the real-time.

Keywords — Voice Activity Detector, Fourier Transform, Deep Neural Network , Convolutional Neural Network, Long Short Term Memory

1 Introduction and Background

Audio signal may contain various of audible sounds like: human voice, music, rain, and another non speech sound. The Automatic Speech Recognition and the other systems that speech processing is only process the sound that contain human speech. That lead to the Voice Activity Detector (VAD) task that decided the input audio contain human speech or not is very important. Basically the VAD systems is processed the input frames $\mathcal{X}$ that is human speech (1) or not (0), a mathematical representation of this problem is:

$$f : \mathcal{X} \rightarrow \mathcal{Y} \quad \text{where} \quad \mathcal{X} \in R^{m \times n} \text{ and } \mathcal{Y} \in \{0, 1\}$$

$m \times n$ is the dimension of input feature that extract from input audio. There are many way to represent the audio signal, let take a look about the theoretical behind them.

Human can hear the sound that have the frequency that between 20 Hz to 20,000 Hz and the Human speech that have the frequency in range 200Hz to 8,000 Hz [1] so by the Nyquist Theorem for lossless digitization the sample rate should be at least twice the maximum frequency response ($f_s \geq 2 * f_m$). So for the human speech, the sample rate is normally 16,000 Hz. Each sample will contain the Amplitude of the signal, that is wave form representation of digital audio.

Sound is a very long signal sequence, but the information content in it is not much, to reduce the storage of sound in time domain, we can represent the sound in the sum of various wave with know the frequency, amplitude and phase of oscillation. Fourier transform help us to do this.

Fourier transform reduce the storage of signal, but still contain the information of signal, we can use Inverse Fourier Transform to convert Frequency domain to time domain. In the recent machine

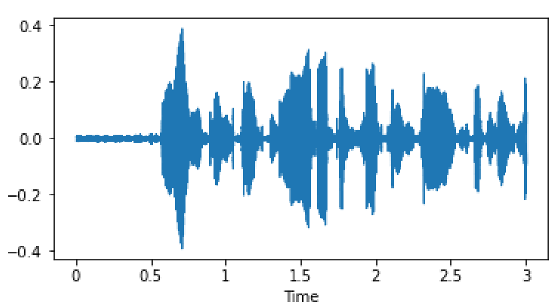


Figure 1: Wave form representation

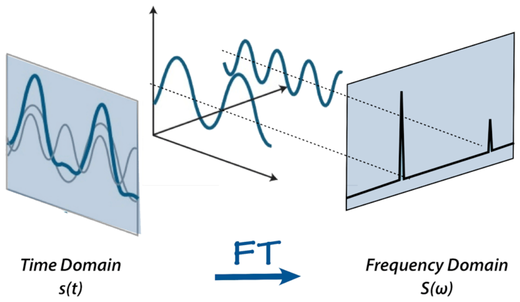


Figure 2: Fouier Transform

learning, deep learning problem, instead of use FT we use the mel-spectrogram to extract the information of audio. Its use DFT with windowing of the frame 20-25ms with overlap length 10ms to capture the context and change context of each frame. Its also use Hamming window, or Hanning window to smooth the frame.

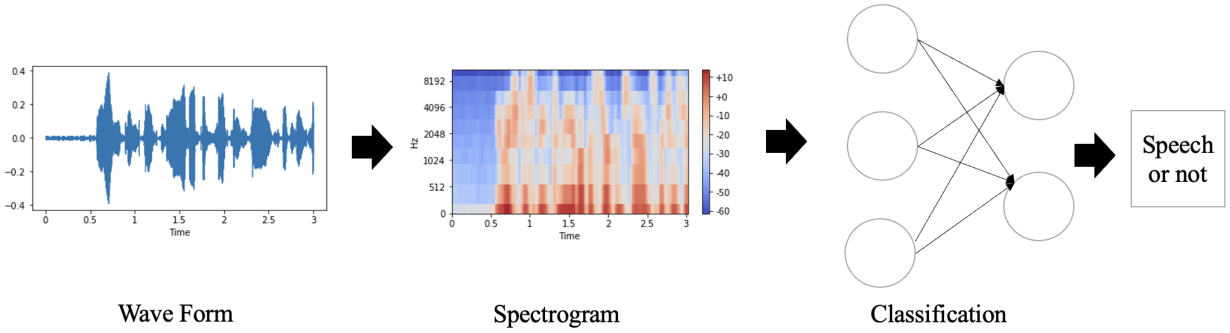


Figure 3: Pipeline in Audio recognition.

This spectrogram will use in the Deep learning task for classification the audio frame is speech or not.

2 Neural Network Architectures for VAD

During many years, there have been many different approaches to solve VAD problem, all with varying degrees of complexity. In this project, I just considering to the Deep Learning method specifically two method that using widely: Long short term memory and Convolutional Neural Network.

2.1 Convolutional Neural Network

Convolutional Neural Net is widely used in image processing and has been very successful. CNN uses conv-layers to extract features of the pixels from which to obtain the image's information. Also techniques like Pooling layers batch normalization layer are used in CNN to imporve the performance When we extract features of Audio using mel-spectrogram, the representation information is image-like, so using CNN is a method worth considering. In 2018 [2] successfully used CNN in VAD in real time on mobile apps achieving 99 % accuracy

2.2 Long short-term memory

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture, LSTM has capable of learning order dependence in sequence prediction problems. According to Wikipedia ”A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well suited for classifying, processing and making predictions based on time series data, as there can be unknown delays between significant events in the time series.”

That leads to LSTM being widely used in audio processing problems and VAD is a typical example. In 2013 Eyben and al [3] successfully used LSTM in predicting VAD and applied it to Hollywood movie dataset, the accuracy reached 97%.
Because of the success of the above methods, in this project I used CNN and LSTM in solving the VAD problem.

3 Experiment and Result

3.1 Dataset

The dataset use in this project is vad, that contain xxx audio file in wav format, it is mono channel, the sample rate is 16000 Hz and the total of time is 3 hours. Each file has labelled the voice of speech in the json file respectively. The figure show the statistic of the dataset.

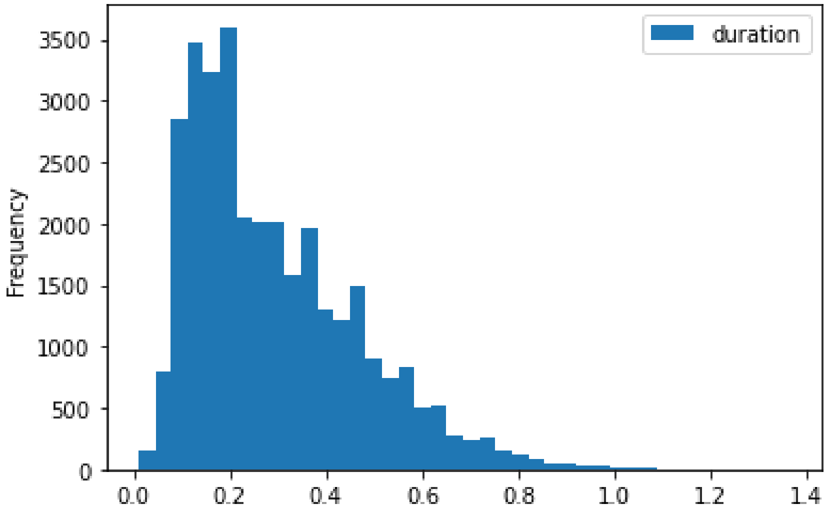


Figure 4: Statistics of VAD dataset

The dataset have total 32560 speech segmentation,the length speech have the range from 0.01s to 1.36s, with the mean is 0.3s. The ratio of speech and total audio time is 0.8. In practice the speech with length < 0.064s (1024 samples) had been through.
Test train split: I will use 70% number of audio for training and the other 30% for testing. The details of train and test file are listed in the source code folder.

3.2 Features extraction

With the dataset and the speech given I will extract the features 1024 samples of each audio by mel-frequency cepstral coefficient, delta, delta order 2 and root mean square energy. The number of MFCCs to return is 5, and length of the Fast FT window is 512 and the hop length is 16 samples. All of these feature will be concatenate each other create a matrix with size 65×16. These sample will label 1 (mean speech), these other samples which not in the labels speech will labeled with 0 (mean noise). After extract all feature of vad dataset I have total **164,068** records.

3.3 Training and results

For training a VAD model, I will implement 4 models: Fully Connected, LSTM, CNN, Resnet1D.

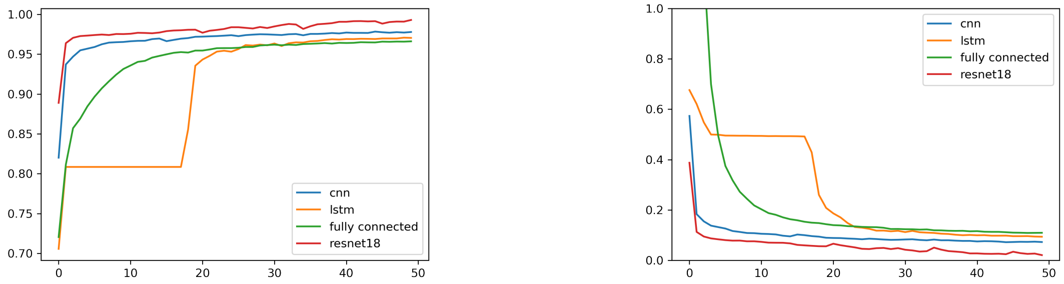
1. Fully Connected Network with the units is 512, 256, 128, 64, I also add the drop out rate to 0.2 for avoiding overfitting.

2. The LSTM model have two layers lstm with 128 head.

3. CNN with Con2D layer of 32, 64, 128 filters, kernel is 3 and 5.

4. ResNet18 1D base ond Resnet18 architecture but the input shape have 1 channel instead of 3 channel, and the output layer have 2 units.

I trained with **50** epochs, using Adam optimizer, the learnig rate is 0.01. Because of the binary classification problem, so the loss ”binary crossentropy” and the metric for the evaluation is accuracy are chosen. The total train time is $\tilde{1}0$ minutes with GPU **NVIDIA RTX A6000** with batch size is **8096**.



(a) Training accuracy of 4 models

(b) Training loss of 4 models

Figure 5: Train accuracy of 4 models in the VAD dataset

As we can see in the Figure 5 the model ResNet have a best performance in the training dataset for both the accuracy and loss. The fully connected model have lowest performance. This is quite understandable because CNN and LSTM models often give better performance when compare with the fully connected model.

Let take a look in the accuracy in the testing set, the table 1 give us the details of each model. Although of Restnet have an accuracy 99.29% in training set but the accuracy in testing set is not well as these other model. LSTM model has the fewest parameters but gives good accuracy in both dataset.

Table 1: Accuracy and number of parameters of 4 models

Network	Params	Train Accuracy	Test Accuracy
Fully-Con	0.7M	0.9661	0.9694
LSTM	0.2M	0.9704	0.9663
CNN	2.0M	0.9778	**0.9746**
Resnet1D	11M	**0.9929**	0.9654

3.4 Training with external dataset

As we know the vad dataset has the ratio between speech and non-speech is 8/2, so I think the adding more data for training is very important. The LibriSpeech dataset has the same as the vad, is a corpus of approximately 1000 hours of 16,000Hz read English speech. For training in this project, I just use the test dataset and the speech is automatic labeled by using a powerful VAD tools. The result is better in training set (about 1%) but the accuracy in testing set is less than about 0.02%. The logs of trainig model can be found in source code.

3.5 Testing in the audio

After training the model, I want to test predict in an audio to verify the performance of model. So I chosen an audio that did not use in train and predict all sub segment samples.

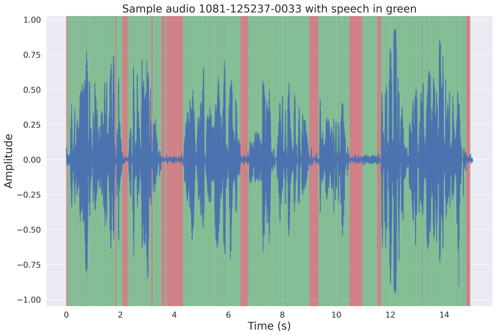


Figure 6: Voice activity Detector in a audio.

As we see, the model predict pretty good in overall. There are some sub samples from the seconds 11 and 12.

3.6 Voice Activity Detector in real-time

Not stopping at prediction in audio files, I want to test the performance of my model in real-time which means I will test in the streaming audio, the basic is a record of the microphone. To make it, I use the python library PyAudio, record the audio from the laptop's microphone with the same sample rate of 16,000 Hz as audio, and the Chunk size is 1024 samples. Each frame is recorded and then predicted by the model. By the calculating, it takes 167ms for extracting the features with Fourier transform and the model takes 64ms for predicting the result. In total, It takes 250ms to the record, extracts and predict, the latency is accepted in the real-time systems.

To make the visualization, I draw the predicted probability for each frame. For the result, the model predicts very well the speech, and the ambiance sound, but some noise sound like the clap

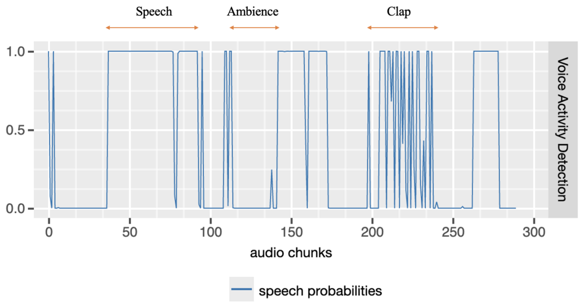


Figure 7: Sample figure caption.

the model predicts as a speech. But when we take a look at the probability, the clap length is very short, and the diagram is not like the real speech. So we can handle this problem by limiting the length of the speech.

4 Conclusion

Overall, this challenging project on VAD was remarkably interesting as it gave me the opportunity to work on a difficult research topic. During this week of research, I conducted several experiments on deep learning for VAD such as LSTM, CNN and also deployment the model for predicting audio and the real-time systems.

Given more time I would continue to improve the performance of the model by using another feature of audio or using some technique of speech processing as smoothing or limiting the length of the speech. Integrating VAD into mobile devices is also an interesting thing to implement.

References

[1] Klara Nahrstedt. *CS 414 – Multimedia Systems Design Lecture 3 – Digital Audio Representation*. 2012. URL: https://courses.engr.illinois.edu/cs414/sp2012/Lectures/lect03-audio-representation.pdf (visited on 02/06/2022).

[2] Abhishek Sehgal and Nasser Kehtarnavaz. "A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection". In: *IEEE Access* 6 (2018), pp. 9017–9026. DOI: 10.1109/ACCESS.2018.2800728.

[3] Florian Eyben et al. "Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 483–487. DOI: 10.1109/ICASSP.2013.6637694.