4. Experiment

4.1 Datasets

4.2 Feature Extractor

4.3 Model

4.4 Training and Testing

4.5 Result

5. Conclusion

4. Experiment:

My practical experiment of the proposed hybrid DNN model is completed only with the first part of the model named vocabnet and it need some background improvement to meet the accuracy as proposed. Eventhough current accuracy lacks with the pre-existing system, changes in the model functions and increasing the training samples with combining the second part will land with a promising result. This section illustrate my works guided by some experienced people working in Deep learning.

All the works are been implemented in python programming languages with the help some module. Following are the module used in this project.

1. Numpy
2. Tensorflow

This project consist of four main phases, the first phases is collecting the sound files which are in wave format and output vocabulary text file and mapping them, further converting the mapped wave file to the MFCC feature vector. The second phase is constructing the model with respective layers, input and output dimension. The third phase is training the constructed model with the dataset collected and reducing the error. And the final phase is to test the trained model with the test dataset.

4.1 Datasets:

This model is trained with two main dataset collection namely TIMIT and VCTK. Where TIMIT can be used for both Vocabulary identification and Phoneme identification but VCTK can be only used for Vocabulary identification. There are many such dataset collection available like \_\_\_\_\_\_\_

4.1.1 TIMIT database

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. Tables 1 shows the number of speakers for the 8 dialect regions, broken down by sex. The percentages are given in parentheses. A speaker’s dialect region is the geographical area of the U.S. where they loved during their childhood years. The geographical areas correspond with recognised dialect regions in U.S. [], with the exception of the western region (DR7) in which dialect boundaries are not known with any confidence and dialect region 8 where the speakers moved. The text material in the TIMIT prompts (found in the file "prompts.doc") consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. The dialect sentences (the SA sentences) were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences (the SX sentences) and each text was spoken by 7 different speakers. The phonetically-diverse sentences (the SI sentences) were selected from existing text sources – the Brown Corpus (Kuchera and Francis, 1967) and the Playwrights Dialog (Hultzen, et al., 1964) - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts. Each speaker read 3 of these sentences, with each sentence being read only by a single speaker. Table 2 summarizes the speech material in TIMIT.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dialect Region** | **Male** | **Female** | **Total** |
| 1 | 31 (63%) | 18 (27%) | 49 (8%) |
| 2 | 71 (70%) | 31 (30%) | 102 (16%) |
| 3 | 79 (67%) | 23 (23%) | 102 (16%) |
| 4 | 69 (69%) | 31 (31%) | 100 (16%) |
| 5 | 62 (63%) | 36 (37%) | 98 (16%) |
| 6 | 30 (65%) | 16 (35%) | 46 (7%) |
| 7 | 74 (74%) | 26 (26%) | 100 (16%) |
| 8 | 22 (67%) | 11 (33%) | 33 (5%) |
| 8 | 438 (70%) | 192 (30%) | 630 (100%) |

Table

VCTK corpus:

This CSTR VCTK Corpus includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph intended to identify the speaker's accent. The newspaper texts were taken from The Herald (Glasgow), with permission from Herald & Times Group. Each speaker reads a different set of the newspaper sentences, where each set was selected using a greedy algorithm designed to maximise the contextual and phonetic coverage. The Rainbow Passage and elicitation paragraph are the same for all speakers. The Rainbow Passage can be found in the International Dialects of English Archive: (http://web.ku.edu/~idea/readings/rainbow.htm). The elicitation paragraph is identical to the one used for the speech accent archive (http://accent.gmu.edu). The details of the the speech accent archive can be found at http://www.ualberta.ca/~aacl2009/PDFs/WeinbergerKunath2009AACL.pdf  
  
All speech data was recorded using an identical recording setup: an omni-directional head-mounted microphone (DPA 4035), 96kHz sampling frequency at 24 bits and in a hemi-anechoic chamber of the University of Edinburgh. All recordings were converted into 16 bits, were downsampled to 48 kHz based on STPK, and were manually end-pointed. This corpus was recorded for the purpose of building HMM-based text-to-speech synthesis systems, especially for speaker-adaptive HMM-based speech synthesis using average voice models trained on multiple speakers and speaker adaptation technologies.

4.1.3 Dataset ordering:

For this experiment I combined both TIMIT and VCTK dataset and cleaned some unnecessary datasets which provide totally \_\_\_\_ sentence. For vocabulary identification only the input wave file and the output character text file is necessary hence I removed all the .PHN and .WRD files from the TIMIT dataset. Then indexed all the input and output file name to the table which further referred by a maaping program that maps all the input and output to a .csv file for feature extractor reference.

4.2 Feature Extractor:

We have dicussed enough about Feature extractor and the need for the feature extractor in Learning system, particularly for speech recognition.

In this section I will will describe the feature extractor system I constructed with MFCC technique. Even though the maxpooling method is been used in the wavenet research [], I find it difficult to map feature for sentence level labels for the model. So, I decided to adopt MFCC feature for this project.