SSP Lab2 Report

Group 3

Member

110061589，110061544，111061612，111064531，111064539，112064535

Motivation

The motivation behind this lab is to first observe the distribution of the test dataset, focusing on various aspects such as data format, gender ratio, and the rules for concatenating audio files. By doing this analysis , it can provide us a clear insights into the distribution of the test dataset. This enables us to generate a corresponding training dataset that is balanced and well-distributed. Our goal is to enhance the diversity of the training dataset through data augmentation, will improve the accuracy of our HMM models.

Test Dataset Check

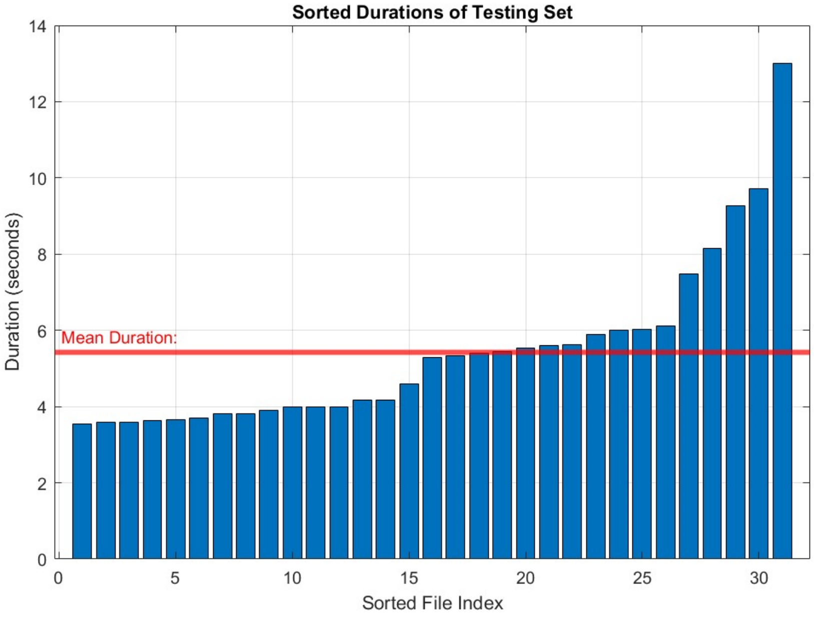
1. Gender Ratio
   - Female=8 people (23%)
   - Male=24 people (77%)
2. Audio Duration

   We visualized the durations of the audio files in the test dataset using statistical methods and discovered that nearly 78% of the files have a duration of less than 6 seconds. However, it is noteworthy that there is one file in the test dataset with an unusually long duration. Upon listening to this particular file, we found that it does not sound like a typical human voice but rather resembles a robotic voice.

   The average duration of audio files=5.42 seconds

   Minimum duration= 3.54 seconds

   Maximum duration = 13 seconds



Make Training Dataset

1.Simply count the occurrence of numbers

In this section, we simplify calculate the probability of the occurrence of each number in the test dataset. The table provides a summary of the occurrences and rates of individual digits from 0 to 9 in a data set.

Total Count: There are 286 total observations.
Count: This row shows the amount of times each number appears in the data set.
Rate: This row shows the percentage of the total count that each number represents.

	total	0	1	2	3	4	5	6	7	8	9
count	286	44	95	34	17	13	25	34	4	6	14
rate	100%	15%	33%	12%	6%	5%	9%	12%	1%	2%	5%

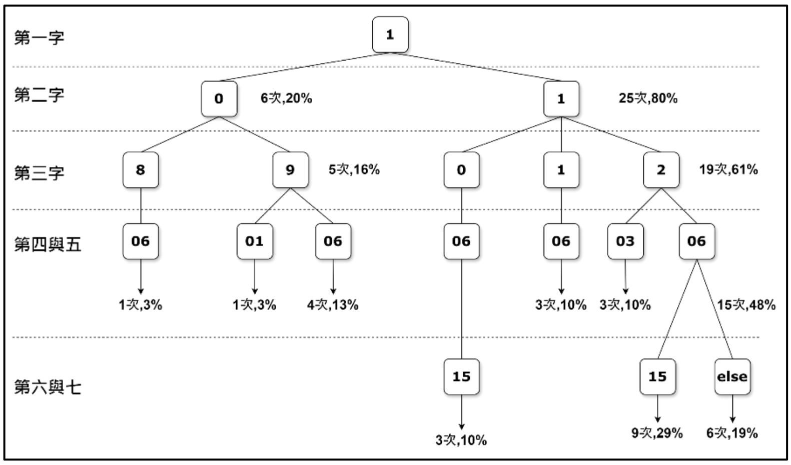
2.Calculate the probability of each bit combination occurrence

Calculate the statistics of all possible digit combinations, taking into account the sequential relationships. The following rules were applied for concatenating the audio files:

1. The first digit must be: 1
2. The second digit:
   - ' 1 ' = 25 (80%)
   - ' 0 ' = 6 (20%)
3. The third digit:
   - ' 0 ' = 3 (10%)
   - ' 1 ' = 3 (10%)
   - ' 2 ' = 19 (61%)
   - ' 8 ' = 1 (3%)
   - ' 9 ' = 5 (16%)
4. The 2 & 3 digit sequence:
   - ' 10 ' = 3 (10%)
   - ' 11 ' = 3 (10%)
   - ' 12 ' = 19 (61%)
   - ' 08 ' = 1 (3%)
   - ' 09 ' = 5 (16%)
5. The 4 & 5 digit sequences:

    - ' 01 ' = 1 (3%)
    - ' 03 ' = 4 (13%)
    - ' 06 ' = 25 (81%)
    - ' 20 ' = 1 (3%)

6. The 1 & 2 & 3 & 4 & 5 digit sequences:
    - ' 11006 ' = 1 (3%)
    - ' 11106 ' = 4 (13%)
    - ' 11203 ' = 3 (10%)
    - ' 11206 ' = 15 (48%)
    - ' 11220 ' = 1 (3%)
    - ' 10806 ' = 1 (3%)
    - ' 10901 ' = 1 (3%)
    - ' 10906 ' = 4 (13%)
7. Digits Occurrence prob:



By following the rules for concatenating audio files, it helps us in creating a balanced and well-distributed dataset, essential for achieving accurate in this tasks.

Data augmentation

First, we converted the concatenated 1,000 audio files into mono channel. Then, we applied three main processing methods: pitch shifting, voice activity detection (VAD), and voice conversion to increase data diversity.

1. Pitch shifting & Speed up

   After initially analyzing the duration of the audio files in the test dataset, we attempted to synthesize data of the same length. However, when the audio was speed up or slowed down, the pitch of the sound changed significantly. When the audio file is sped up (making the file duration shorter), the sound frequency becomes higher. Conversely, when the audio file is slowed down (making the file duration longer), the sound frequency becomes lower. Therefore, we used pitch shifting to adjust the speed-up or slowed-down audio files to sound more like the original audio.

   The relation of speed up ratio & pitch:

   $$Pitch_{adjust} = -12 * log_2(speed\ ratio) \quad , speed\ ratio > 0$$

   $$Pitch_{adjust} = 12 * log_2(speed\ ratio) \quad , speed\ ratio < 0$$

   When the speed ratio is greater than 0, the pitch needs to be adjusted downwards. Conversely, when the speed ratio is less than 0, the pitch needs to be adjusted upwards."

2. voice activity detection

The process involves segmenting the audio into frames of 10 ms and applying a padding duration of 10 ms to effectively remove silence

3. voice conversion

The target speaker ID '0084' (male, age 20-24) will be applied to each audio file, resulting in a new set of 1000 voice-converted audio files

Training Dataset Combination

Data Assemble number refers to the method in “Make Training Dataset”

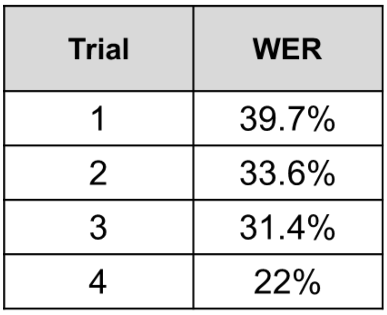
All digit combinations use a male-to-female ratio of 2:1

Data Assemble	Digits length	Speed up & pitch	Remove Silence	Voice Conversion	Data Number
1	9	no	no	no	1000
2	9	Yes	no	no	1000
2	9	Yes	Yes	no	1000
2	9	Yes	Yes	Yes	1000

Training process

we had 1000 audio files in the dataset. By applying the three methods (pitch shifting, voice activity detection, and voice conversion), we added 1000 new audio files for each method. This resulted in a total of 4000 audio files in the dataset.
And for the training process, the parameter we set:
Number of iterations:160
Last iteration to increase the number of Gaussians:120
Target number of Gaussians:300

Result

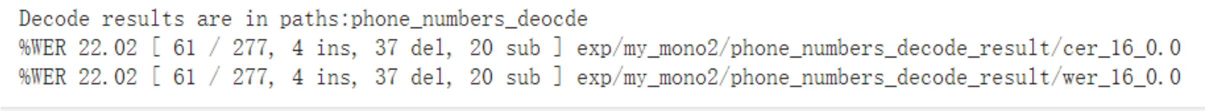


Trial	WER
1	39.7%
2	33.6%
3	31.4%
4	22%

Despite applying a tri-phone model, the performance did not surpass the mono-phones model. Considering our training data, involving Chinese numbers from 1 to 10, the mono-phones perform better. Here is some reasons we thought:

1. Small Data Size**:** The phoneme set for Chinese numbers from 1 to 10 is limited, resulting in a relatively small dataset. In such cases, monophone models, with fewer parameters and lower complexity, might achieve better performance.
2. Context Dependency: The pronunciations of Chinese numbers are relatively independent and stable. This makes mono-phone model more effective.

Our training results show that the mono-phones model performs better under these conditions.
Best Model Performance:



```
Decode results are in paths:phone_numbers_deocde
%WER 22.02 [ 61 / 277, 4 ins, 37 del, 20 sub ] exp/my_mono2/phone_numbers_decode_result/cer_16_0.0
%WER 22.02 [ 61 / 277, 4 ins, 37 del, 20 sub ] exp/my_mono2/phone_numbers_decode_result/wer_16_0.0
```