ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

Applications of Machine Learning on Translation between Digits

Ziyi Zhuang[1,*], Minghao Sun[2], Qinghan Li[3]

[1]YanDao road High School, Chengdu, Sichuan 610016, China;

[2]Guangzhou University Affiliated Middle School, Guangzhou, Guangdong, 510000, China;

[3]Shenzhen college of international education, Shenzhen, 518043, China.

*Corresponding author. Email: juliannnaguan@gmail.com

Abstract

In order to realize digit speech translation, system was researched by template matching and artificial neural network through the software programming. The applications of digit translation system were studied, and the Chinese and English digital translation system was constructed in Matlab. Experimental results show that different algorithms reach different accuracy.

Keywords

Speech Recognition; Digital Translation System; TM; Matlab; BNN; BNN2HL; Neural Network.

1. Introduction

Early translation of the text and the speech is mainly finished by persons. Nowadays, Computer technique, Machine Learning [1] and digital audio signal processing [2,3] provide good implementation scheme of the machine translation both on the text and the speech. There are many product or technique solution for machine translation in the world, for example, Google translate [4] is the technique of Google's free service instantly translates words, phrases, and web pages between English and over 100 other languages. Iflyrek translation [5] is the on-line solution of the speech translation between Chinese and several mostly used languages, such as English, Spanish and Russian.

In this report, Deep Neural Network [6], particularly Template Matching,Binary Neural Network and Artificial Neural Network are used to fulfill Digit Speech Translation between Chinese and English.

2. Digit Speech Translation between Chinese and English Based on different method

2.1 Template Matching [7]

Template matching (TM), which is one of the oldest and widely applied Machine Learning (ML) methods, is relatively simple using the important sum-of-products operation: Given template row vector T containing n elements Ti for $1 \leq i \leq n$ and data feature row vector x containing n elements xi, the output y measures the similarity between the T and x vectors computed as the sum-of-products as

$$\mathbf{y} = \sum_{i=1}^{n} Ti\ xi \quad (1)$$

Figure 1 shows the block diagram of a simple TM structure

For digit classification, each digit is represented by its own spectral or MFCC feature template; for example T0, T1,..., and T9 ; and each x vector is applied to ten similarity computations to produce outputs y0,y1,...,y9. The digit is then classified by the maximum y# that is observed.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

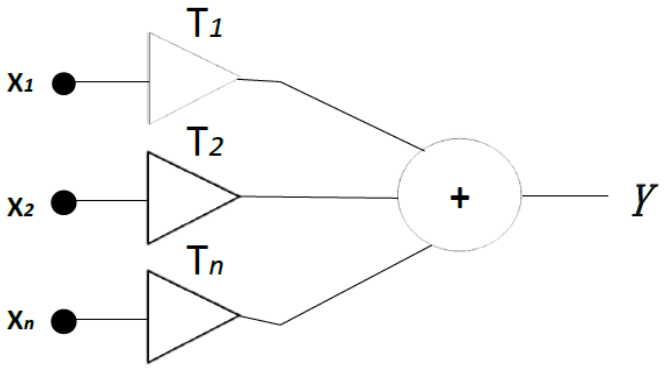

Figure 1. Simple TM structure

TM comes in two flavors, depending on whether the template is known or whether it must be estimated from the data.

2.2 Mel Frequency Cepstral Coefficients (MFCC)

Mel-scale Frequency Cepstral Coefficients (MFCC) is cepstral coefficients that extracted from Mel-standard-frequency, which describes what human hear has non-linear characteristic.

The Cepstrum of speech signal is given as

$$C(q) = IFT(S_{dB}) = IFT(H_{dB}(f)) + IFT(G_{dB}(f)) \quad (2)$$

The relation-ship between Mel-standard and frequency can be described as the following formula where f is the frequency in Hz.

$$M(f) = 1125 \ln(1 + f/700) \quad (3)$$

The discrete cosine transform (DCT) of an n-element vector of real values, the log-Energy of PSD values, produces an n-element vector of real values. Because the input vector corresponds to frequency values, the DCT produces a vector of time-like values. Because of the nonlinear logarithmic operation, the units are not seconds but frequency, a play on the word frequency, and the result is called the cepstrum, a play on the work spectrum. Because we used a Mel filterbank, a set of Mel-filtered cepstral coefficient (MFCC) vector is produced

The process of extracting speech feature based on MFCC is given as following flow chart.

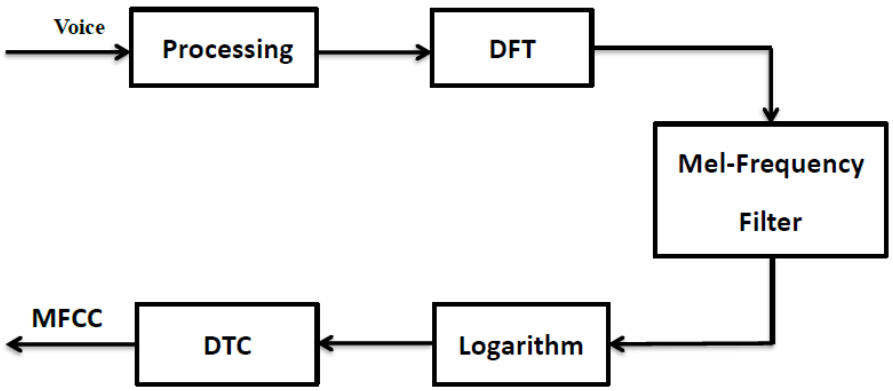

Figure 2. The process of extracting speech feature based on MFCC

Flow chart shows the process that generates MFCC. Because the mean value of the log-Energy has been subtracted, the first element is zero, and why it contains no information. This is the reason the MFCC Tutorial ignores this first element, and it suggests using elements 2 to 13 as the feature vector that contains 13 features per frame times 25 frames, or 325 features, displayed as an image.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

2.3 Binary Neural Network (BNN)

In *logistic regression* the ANN output *y* values are either 0 or 1, which are usually limited to only two-class classification problems implemented with *binary NNs*, denoted *BNN*.

The BNN output value is computed as [e-z]

$$y = \sigma(\sum_{i=1}^{n} WiXi) \quad (4)$$

where the weight vector W with *n*-elements Wi replaces the template vector *T*.

The $\sigma(z)$ is the nonlinear *logistic function* computed as

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (5)$$

As $z \rightarrow -\infty$, $e^{-z} \rightarrow \infty$, $\sigma(z) \rightarrow 0$.

As $z \rightarrow \infty$, $e^{-z} \rightarrow 0$, $\sigma(z) \rightarrow 1$.

Figure.3 shows the block diagram of a binary NN

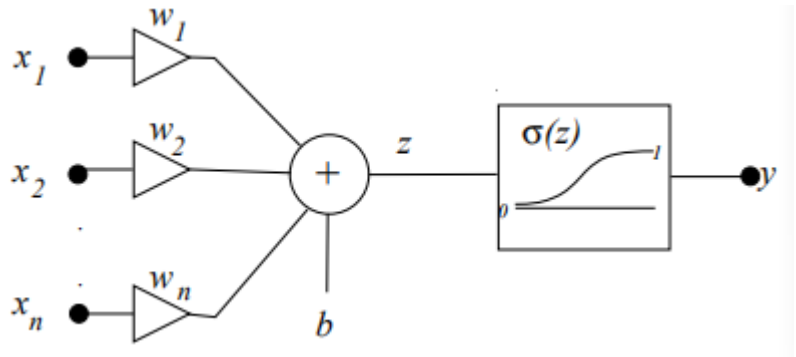

Figure 3. Binary NN architecture that classifies one of two objects from an n element feature vector

```
SI vector for random spoken digit 9
     0       1       2       3       4       5       6       7       8
9
 0.267  0.165   0.062   -0.082  0.059   0.364   0.296   0.064    0.030   0.
611   Speech classification by maximum similarity index gives 9
```

Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	0	4	0	0	0	0	0	7	0	0
1	0	0	0	0	0	0	0	8	0	0
2	0	0	0	0	0	0	0	0	9	0
3	0	0	1	3	0	0	0	0	7	0
4	0	0	0	0	6	0	0	0	0	0
5	0	0	0	0	0	8	0	0	0	2
6	0	0	0	0	0	0	10	1	0	2
7	0	0	0	0	0	0	0	10	0	0
8	0	0	0	0	0	0	0	0	9	0
9	0	0	0	0	0	0	0	0	0	13

Trials= 100 Errors= 41 P_err= 0.410

Figure 4. Confusion Matrix and PE of the digit speech Translation from Chinese to English

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

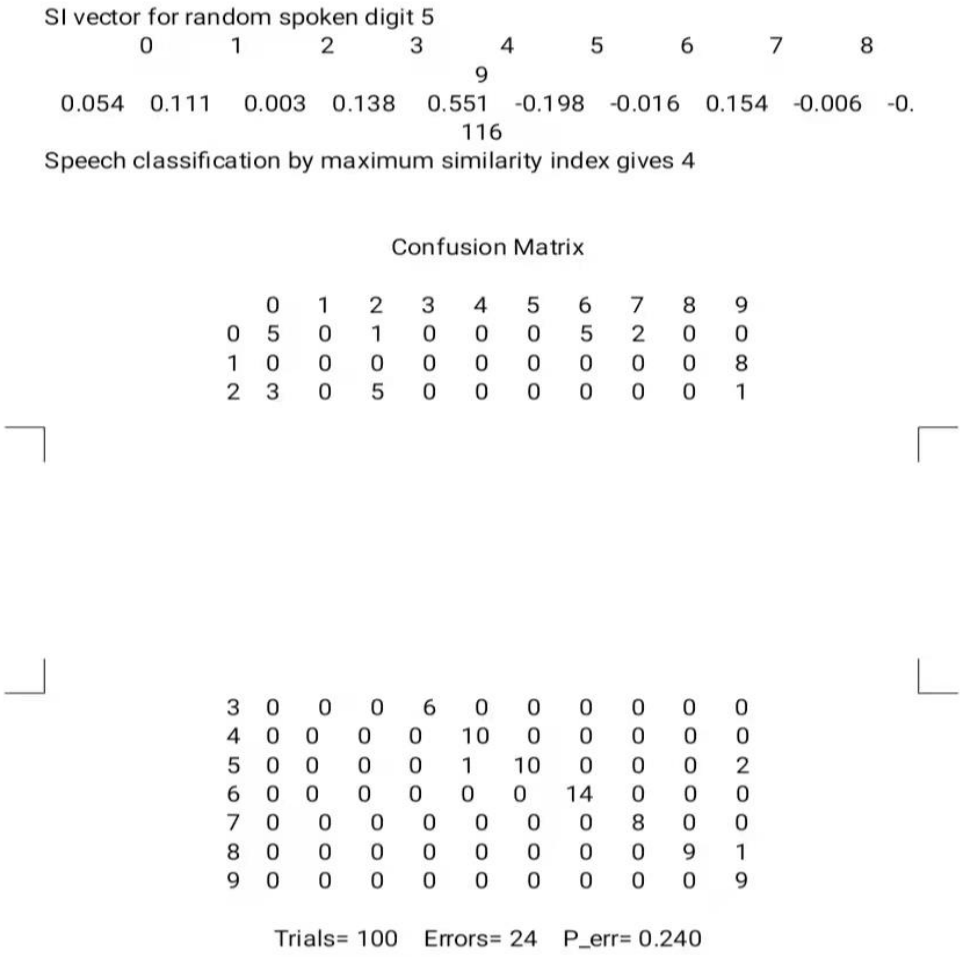

Figure 5. Confusion Matrix and PE of the digit speech Translation from English to Chinese

3. Experiment

3.1 Template Matching Program

Two Matlab programs, TM_RT_TranslationCtoE.m and TM_RT_TranslationEtoC.m, are respectively written to realize the digit speech Translation between Chinese and English and the digit speech Translation between English and Chinese. Two digit speech databases, S9C.mat and S9E.mat, are applied to the template matching. And digit speech, that is, in the digit speech Translation from Chinese to English; the input is the Chinese digit speech, the templates is from S9C.mat and the English digit output is based on S9E.mat and max SI of Chinese Digit classification; in the digit speech Translation from English to Chinese, the input is the Chinese digit speech, and the templates is from S9E.mat. The English digit output is based on S9C.mat and max SI of English Digit classification.

The Confusion matrix, PE of the digit speech Translation from Chinese to English, and the digit speech Translation from English to Chinese are respectively shown in Figure.4 and Figure.5

3.2 Binary Neural Network Program

Firstly, the language recognition system is carried out. The digital speech used in the experiment is collected in a quiet environment. The 20 digital speech signals of "0-9" in Chinese and English are feature extracted, and then training and recognition are started.

The stage of training: Bnn2hl Train program will be used to train the previously collected voice signals, and BNN2HL Weights will be obtained for real-time language recognition.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

Real-time identification stage: The real-time speech phase is divided into two parts, the first part is language recognition (LR), and the second part is digit recognition (DR). Jeremymain is used to determine which database the DR program uses. In order to make the user s operation easier, we have stored the real-time input voice signal in the LR program, so as to avoid the user s second input of voice signal.

Voice output stage: We downloaded the Chinese and English 0-9 voice respectively from the Internet. In the Dr Program, the voice number will be output, and then according to the voice number, the corresponding voice will be output, so as to achieve the role of digital translation.

Figure.6 (down) describes the entire process.

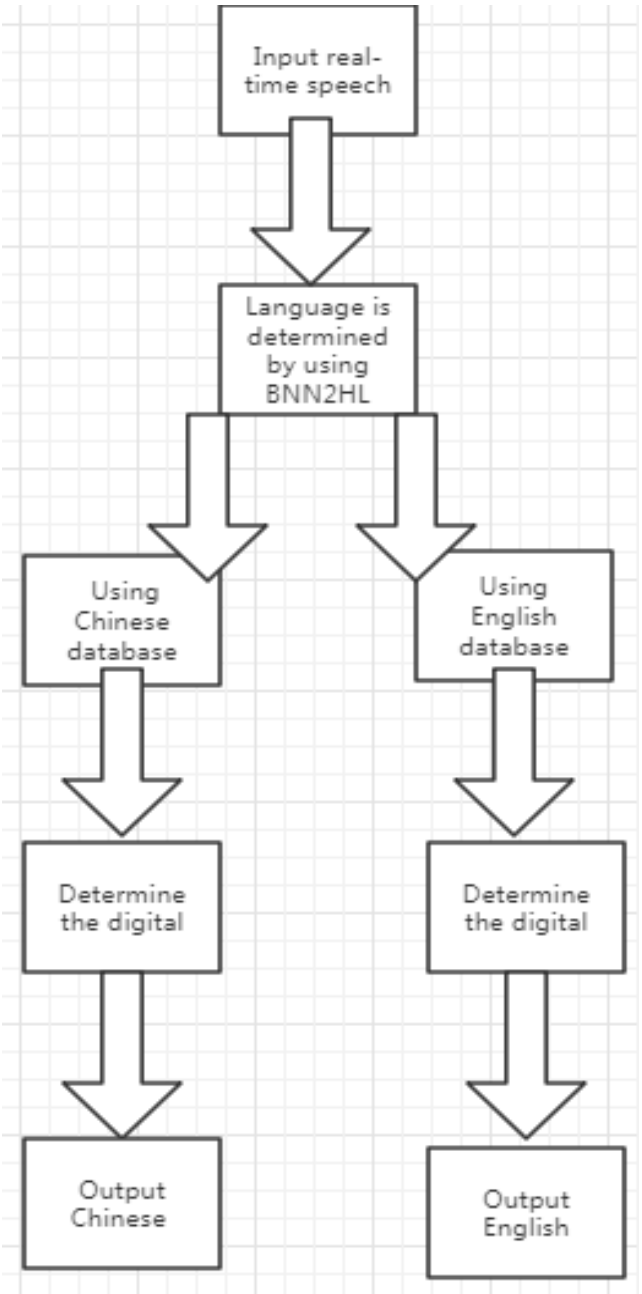

Figure 6. Flow chart of binary neural network system

The Confusion matrix, PE of the digit speech Translation from Chinese to English, the digit speech Translation from English to Chinese and the recognition of binary neural network are respectively shown below.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

BNN2HL Language Recognition
Confusion Matrix
0 1
0 52 0
1 2 46
Trials= 100 Errors=2 P_err= 0.020

Chinese
Digit Recognition
Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	8	1	0	0	0	0	0	0	0	0
1	1	11	1	0	0	0	1	1	0	0
2	0	0	8	4	0	0	0	0	4	1
3	0	0	0	2	0	1	0	0	0	0
4	0	0	0	1	0	1	0	0	3	0
5	0	1	0	0	0	9	0	0	0	0
6	0	0	0	0	0	1	1	0	2	0
7	2	5	0	0	0	0	0	8	0	0
8	0	0	3	1	0	0	0	0	8	0
9	0	0	0	1	0	0	0	0	1	9

Total # of test vectors= 100 # of errors= 36 P_err= 0.36

English
Digit Recognition
Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	3	0	2	0	1	2	0	0	1	1
1	0	4	2	0	1	2	0	1	0	0
2	2	1	7	1	1	1	0	0	0	0
3	1	0	0	5	0	0	0	0	1	0
4	2	0	0	0	7	4	0	0	0	0
5	0	0	2	0	2	4	0	2	1	0
6	0	1	1	0	0	0	1	0	6	1
7	1	0	3	0	1	0	0	3	0	0
8	1	0	3	0	0	0	0	0	4	0
9	0	0	0	0	4	1	0	0	3	2

Total # of test vectors= 100 # of errors= 60 PEG= 0.600

Figure 7. Confusion Matrix and Percentage of Error of BNN2HL

3.3 Digital Neural Network Program

This program mainly uses digital ANN with two hidden layers (DNN2HL) as the only way to classify those input data. When compares to linear digital ANN, the DNN2HL can form a more complex function in order to have a better performance.

Originally, I had thought about other methods to do the parts of classifying language and digits, however, after discussing with my partners, I found that in the traditional process, we should firstly classify the language of the digits, then classify the digits with its corresponding weights, this trend may result in a more complex process, as it need another neural network to deal with the languages, so I would like to classify both language and digits in one step.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

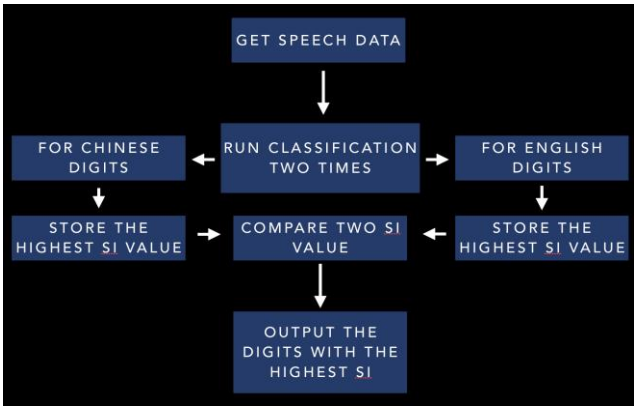

Figure 8. Flow chart of digital neural network program

the whole program is quite straight forward, as every time when classify a digit, the program will calculate the value of similarity each time, in this situation, I can simply compare the maximum similarity from both classification for Chinese and English, Where the digit with highest similarity should be the corresponding digit that the user input.

I Justify the program to store the speech data into an array in this first time, then run the classification step one by one, each time the program will store an array with two values.

In detail, the program gets three parts, main part, classification part, and also the output part. The main part is used to call other two functions. Also, the main part can be used to figure out the maximum value of similarity from both two languages.

Firstly, the main part program call the function of classification twice with different parameters, “E” and “C”, those two parameters will let the classification parts to give similarity values from the input signal with the weights of both Chinese and English data. Then every time the function will return two values, both of which are stored in a vector, the similarity value and the corresponding value.

Then, the main part of program will compare the similarity value from both two vectors, which are from classification for English digits and from the Chinese digits. According to the act that, if the input is 7 in English, in this situation, the similarity from the classification from English digits will be significantly higher than the result from classification of Chinese digits. So that this part can figure out the which digits it should belongs to.

```
the input language is English
the Confusion matrix for Chinese digit:    9
          0     1     2     3     4     5     6     7     8     9     0     1     2     3     4     5     6     7     8     9
    ------------------------------------------------------------------------------------------------------------------------

    0||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    1||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    2||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     2     0     0     0     0
    3||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    4||    0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     2     0     0
    5||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    6||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    7||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    8||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    9||    0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0
    0||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    1||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    2||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    3||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    4||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    5||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    6||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    7||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    8||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
    9||    0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0
PEG =
     1
```

Figure 9. Confusion Matrix and Percentage of Error of DNN2HL(input 9)

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

In the end, the output part function will play the pre-recorded speech file to output sound in order to achieve the aim of translation. There will be two parameters, the 'Lan' for language and the 'n' for corresponding number, those will be used for the last function to determine with file should it play.

In addition, in order to test the performance of the whole program, I use a confusion matrix to help me to analysis the error made by this classification.

The first is the corresponding number and the second is it's SI value, in the third part, those two value will be compared and the one with highest similarity will be output.

the Confusion matrix for Chinese digit: 6

	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
0||	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4	0
1||	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
2||	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0
3||	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
4||	0	0	0	0	0	0	0	5	0	0	0	0	1	0	0	0	0	0	0	0
5||	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6||	0	3	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
7||	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0
8||	0	0	0	0	0	0	0	0	1	0	0	0	0	0	4	0	0	0	0	0
9||	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
0||	0	1	0	0	0	0	2	0	0	0	1	0	1	0	0	0	0	0	0	0
1||	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1||	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2||	0	0	0	0	0	0	0	4	0	0	0	0	2	0	0	0	0	0	0	0
3||	0	0	0	0	0	0	0	1	0	0	0	0	0	5	0	0	0	0	0	0
4||	0	0	0	0	0	4	0	0	0	0	0	0	3	0	0	0	0	0	0	0
5||	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	3	0	0	0	0
6||	0	0	0	0	0	0	0	1	0	0	0	0	0	5	0	0	0	0	0	0
7||	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
8||	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0
9||	0	0	0	0	0	1	1	0	0	0	1	0	0	0	4	1	0	0	0	0

PEG =

0.7200

Figure 10. Confusion Matrix and Percentage of Error of DNN2HL (input 10)

However, even the program can barely work normally, the result is unsatisfactory, as the PEG value is abnormally high, and also the confusion matrix seems to be quite random, as the error involved in this program shows a uniform distribution.

After finding the problems which cause the extreme low accuracy, the first factor may be the design of program itself, I believe that the corresponding digits will get the highest similarity, however, this assumption may base on a wrong theoretical basis. That the similarities obtained from both programs may be close to each other, as there may be some similar sound in both Chinese and English, that the program may choose the wrong one when comparing the similarities.

On the other hand, the database of this program is not pre-selected, as there are many training data is low in quality, the training based on that database may not be efficient as well as my expect.

Finally, the quality of microphone may also cause some negative effect on the accuracy, Microphones tend to involve some noise when recording sounds, as well as there are some breakages commonly

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

happened during the recording, those interference will significantly lower the accuracy of this program.

4. Conclusion

4.1 Template Matching

The digit speech Translation between Chinese and English based on Template Matching is accomplished, which include the theory analysis and Matlab programming for the forming of the database for digit speech in Chinese, English MFCC, the forming of the digit template, the template matching, the exchange of speech database, and the translated speech output. Ideal confusion matrix and PE are achieved in the classification and the translation of digit speech.

The future work focuses on the Chinese to English or English to Chinese translation of digit speech based on the conversion of the templates with the help of the multi-layer convolutional neural network, which can be work much usefully. Because though this experiment, the Chinese to English translation shows a lot problems. For example, 2 or 3 was often seen as 8, 1 was often seen as 7, and 0 was never seen right, some of which can be explain-like 1 and 7. These two numbers have similar pronunciation in Chinese. But other the error that happened in identifies of other digits cannot be solved or explained now. Consequently, both of improvements of currently used programs and the introduction of new Neural-network are crucial.

4.2 Binary Neural Network

BP neural network is a hotspot in speech recognition. Due to the limited level and time, the content of this experiment is inevitably deficient, and there is still a great room for improvement. The phonetics database of this experiment contains few number elements, so a larger phonetics database is needed in practical application, which includes not only numbers but also words; therefore the difficulty is much higher than previous. Through this experiment, we have a further understanding of the working principle of BP neural network. It can use neural network to realize some simple recognition and lay a foundation for the application of neural network in the future. The experiment of speech recognition has yet to be continuity, sample data of neural network is extracted from the original voice again after actual speech segment of multiple MFCC parameters derived from the data segment is combined, but the voice is continuous change, so the future can make speech recognition has the continuity, the speech in time domain on the speech recognition system continuously for preprocessing of voice feature extraction.

4.3 Digital Neutral Network

The scheme of finding the maximum similarity by two times of classification can be effective to some extent, but the practical test shows that the accuracy of this method is not high enough. Although the result was not idea, this failure can inspire me about some other solutions to this problem.

If I will make some contribution to improve the performance of this method, it may be necessary for me to use better weight data, as the data base is not enough for a reliable neural network training, also the raw data should be improved as the quality of raw data will dramatically affect the performance of the weights after training, the low quality sound may lead to a wrong weight which will lead to a unexpected low accuracy.

In a nut shell, the results showed that this was not a successful attempt, as this bold approach could not achieve the results I expected.

Acknowledgments

Ziyi Zhuang, Minghao Sun and Qinghan Li contributed equally to this work.

Thank for professor Roman Kuc rendered many ideas during the development of these translation programs.

Thank for the help that Torhea Education Group offered during this course.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0051

References

[1] Roman Kuc, Speech Recognition Using Machine Learning, CIS'2020 Lectures, 2020.

[2] Udo Zolzer, Digital Audio Signal Processing, Wiley & son Ltd. 1997.

[3] Hongsong li, Dongmei Chen, Digital video and Audio Technique,Tsinghua University Press, 2011

[4] Google Translate (Google's free service instantly translates words, phrases, and web pages between English and over 100 other languages), translate.google.com.

[5] Iflytech Speech Translate, www.iflyrek.com

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury,Deep Neural Network For Acoustic Modeling in Speech Recognition,IEEE SIGNAL PROCESSING MAGAZINE [82],2012.

[7] 1. R. Brunelli. Template Matching Techniques in Computer Vision: Theory and Practice, Wiley, ISBN 978-0-470-51706-2.2009 (http:/eu.wiley.com/WileyCDA/WileyTite/productCd-0470517069.htm)