International Core Journal of Engineering Volume 7 Issue 1, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

Machine Classification of Spoken Digits

Wei He[1], Ziqian Deng[2], Jingyu Li[3]

[1]School of Electrical Engineering, Chongqing University, Chongqing,400044, China;

[2]University Of Science & Technology Beijing, Beijing,100083, China;

[3]No.58 high school, Qingdao, 266000, China.

Abstract

This paper uses artificial neural network (ANN) to classify spoken digits based on binary feature sets. Multiple feature sets including cepstral coefficients, spectrogram, adaptive spectrogram and their combinations are applied to find which one minimizes the classification errors. The result is the combination of adaptive spectrogram and cepstral coefficients has minimum classification errors. In the process of training ANN, by setting adaptive spectrogram and cepstral coefficients as input vectors and choosing 2 hidden layers, overtraining is overcome.

Keywords

Classification, Spoken Digit, Binary, ANN, Overtraining.

1. Introduction

Currently, cepstral coefficients[1] and spectrogram[2] are both commonly used as feature sets to classify spoken digits. Spectrogram is the result of short-term Fourier transform (STFT) and adaptive STFT(ASTFT) is proposed in this paper. Moreover, since binary feature sets can accelerate computation speed, this paper implements an artificial neural network (ANN) [3]which is trained with binary feature set or feature combination. Firstly, template matching[4] is applied to determine which feature set has best classification performance. Secondly, the best feature set are chosen to be the input vector of ANN. As the over-training happens in the training and the binary systems, using an ANN with two-hidden layer can reduce the over-fitting.

2. Speech data acquisition

The speech data used in this paper is contributed by 11 people and saved as MATLAB file. Everyone speaks a set of digits, ranging from 0 to 9, for 5 times in English and 5 times in Chinese.

Figure 1 shows an example of the speech waveform for English digit 0.

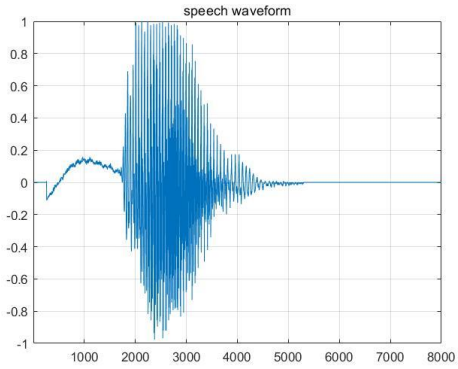


Figure 1. A speech waveform for English digit 0.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

3. Speech processing

There are three kinds of feature sets used in this paper which are cepstral coefficients (CCs), short-term Fourier transform (STFT) and adaptive short-term Fourier transform (ASTFT). ASTFT is based on STFT. It applies a threshold τ to the start and end of the speech waveform to determine the speech size, which is different for each speech signal (adaptive part).

Figure.2 shows the spectrogram of digit 0 before and after applying a threshold.

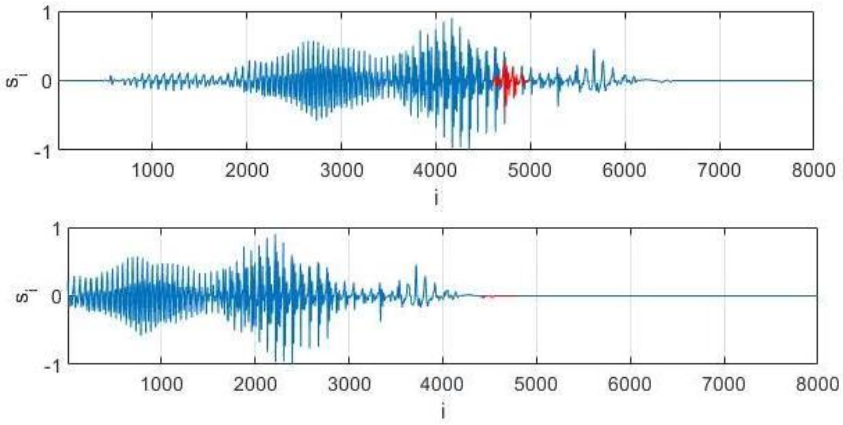


Figure 2. spectrogram and adaptive spectrogram

The range of feature sets is shown in Table.1.

Table 1. The range of different feature sets

	CCs	STFT	ASTFT
range	[-1, 1]	[0, 1]	[0, 1]

The binary function is based on median value of every feature set. For feature combinations such as STFT+CCs, binary process of STFT and CCs are finished respectively before combining them.

Another possible way is to apply a given threshold value such as 0.5, which converts every value in a feature set less than 0.5 to be 0 and larger than 0.5 to be 1. After experiment shown in Table.2, median binary function has better performance.

Table 2. Comparison of two binary functions

	CCs	STFT	STFT+CCs	ASTFT	ASTFT+CCs
THRESHOLD BINARY	66.2%	23.5%	17.8%	17.3%	14.9%
MEDIAN BINARY	48.5%	20.5%	15.6%	13.6%	12.4%

4. Speech classification approach

4.1 Template matching

Template matching is a common method to determine the similarity of two sequences of identical length, the template T_i and an observed data sequence X_i. The similarity index (SI) of T_i and X_i is computed as

$$SI = \sum_{i=1}^{n} \frac{T_i}{\sqrt{\sum_{i=1}^{n} T_i^2}} \frac{X_i}{\sqrt{\sum_{i=1}^{n} X_i^2}}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

Obviously, the range of SI is between -1 and 1. The larger SI is, the more similar two sequences are.

4.2 Neural network

For every digit, execute the following steps while taking CCs as example:

1) calculate its CC signal

2) apply binary function to CC signal

3) save them as TRAIN and TEST databases equally

Thus, the structure of CCs database is shown in Figure.3.

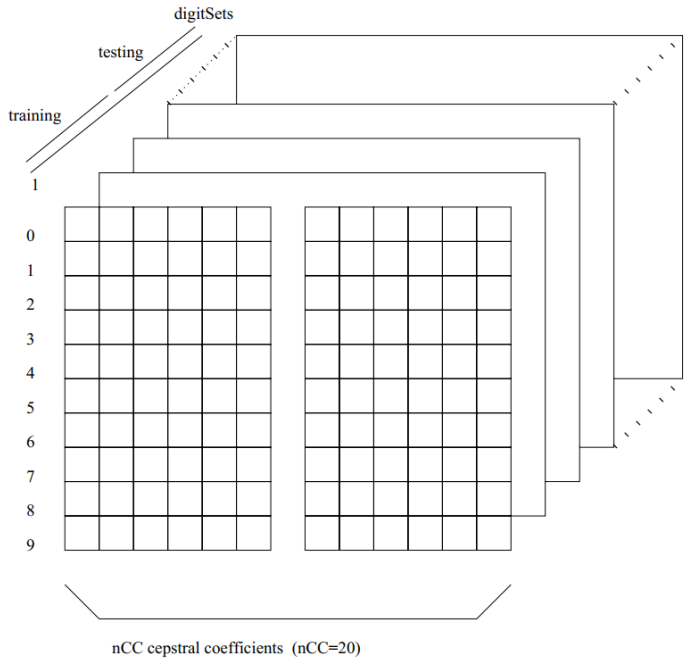


Figure 3. CCs database structure

For other feature sets, the process remains the same.

For one signal (A) in TEST database and another signal (B) in TRAIN database, the closest signal among TRAIN database determines which digit A is classified. Originally, Distance is the sum of squared difference between A and B.

$$\text{Distance} = \sum_{i=1}^{n} (A_i - B_i)^2 \tag{1}$$

where n is the number of features, 30 for CCs and 100 for both STFT and ASTFT. After binarization of A and B, Distance is re-defined as Equation 2.

$$\text{Distance} = sum(bitxor(A, B)) \tag{2}$$

A sigmoid function was also used as Equation 3.

$$y = \frac{1}{1+\exp(h)} = S(h) \tag{3}$$

where the decision variable was given by

$$h = w * x + b \tag{4}$$

$$b = b + -eta * (y - t) * y * (1 - y) \tag{5}$$

$$w = \mathrm{x} * \left(-eta * (y - t) * y * (1 - y)\right) + w \tag{6}$$

where w is the weights between MT and LIP neurons and b is random neuronal noise that cannot be reduced by learning.

Figure.4 shows the model for one hidden layer training, with input x, and output AN_o.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

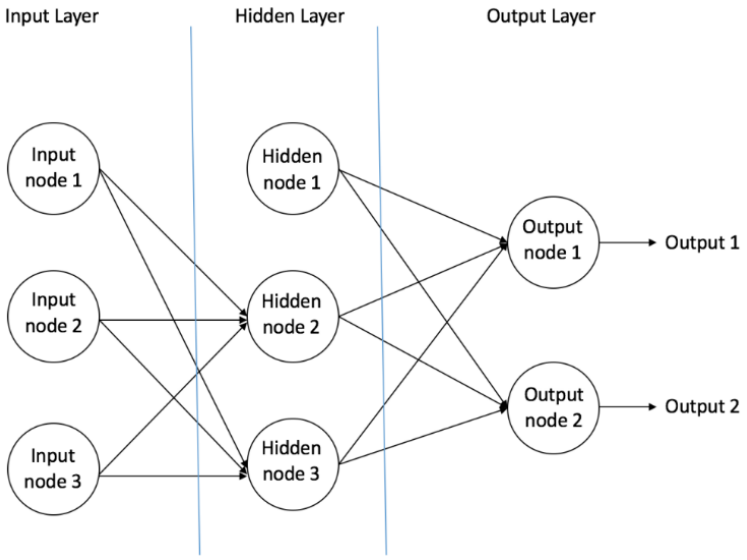


Figure 4. Model for one hidden layer training

In contrast, in the back-propagation algorithm, the output error starts from the output layer and moves backward until it reaches the right next hidden layer to the input layer. Figure.5 shows the training for a two hidden layer training, with four input and three output.

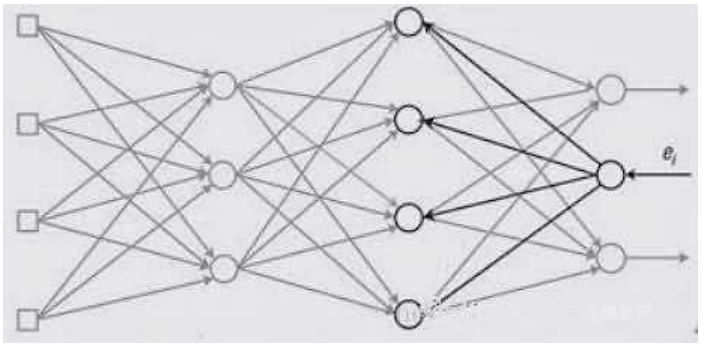


Figure 5. Model for two hidden layer training

Overtraining occurs when the learning rate is too large, causing the boundry is too complicated which represents local feature rather than global feature.

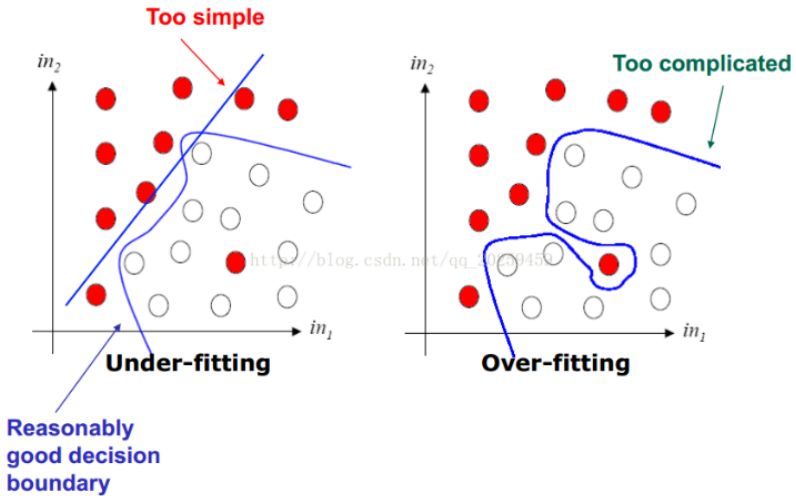


Figure 6. Overfitting in the overtraining

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

4.3 Multiple threshold

When classifying samples of sound, we could apply a threshold to help classification. However, the imporvement in accuracy is not oibvious through one threshold. As a result, a brandnew method is put in use----rather than set one threshold, muptiple thresholds are put in use. Furthermore, the effects of more decimals and normilazaiotn are also combined to contribute to a better effect.

Figure. 7 is an example to set multiple thresholds.

```
if newsignal1(index) > 0.9
    newsignal1(index) = 5;
elseif newsignal1(index) > 0.6
    newsignal1(index) = 4;
elseif newsignal1(index) > 0.5
    newsignal1(index) = 3;
elseif newsignal1(index) > 0.2
    newsignal1(index) = 2;
elseif newsignal1(index) > 0.1
    newsignal1(index) = 1;
else
    newsignal1(index) = 0;
end
```

Figure 7. One example of multiple threshold

Normalization is a way to process data so that the data could be more consistent with independently identically distribution and reduce the deviation caused by internal Convariate shift

Applying more decimal places is to make the threshold more accurate. For example, rather than set 0.2 as the threshold, we can set 0.24 as the threshold, which could lead to a better classification result.

5. Results

5.1 Results of template matching

Figure.8 is the result of STFT using template matching. The row indexes are true digits and column indexes are the results that true digits are classified as. Diagonal elements are the times that digits were classified correctly and off-diagonal elements are errors.

```
E_MS_DB_TM Confusion Matrix
              +-----0----1---2---3---4----5----6---7---8---9--
            0|  44    1    0    1    2    1    2    2    2    0
            1|   0   43    0    0    2    7    0    0    0    3
            2|   1    0   47    2    2    0    2    1    0    0
            3|   1    0    2   48    0    0    2    2    0    0
            4|   0    2    0    0   53    0    0    0    0    0
            5|   0    3    0    0    0   47    0    1    0    4
            6|   0    0    1    0    0    0   51    0    3    0
            7|   3    0    1    0    1    0    0   49    1    0
            8|   0    0    0    1    0    0    3    0   51    0
            9|   1    3    0    0    0    4    0    0    0   47
   E_MS_DB_TM: nTrials= 550  nErr= 70   Prob[err]= 0.127  550 nErr= 222
                             Prob[err]= 0.404
```

Figure 8. The result of STFT

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

It can be found that with a threshold $\tau = 0.15$, the probability of error decreases from 0.404 to 0.124. Fig.8 displays the result of ASTFT+CCs.

```
E_MS_DB_TM Confusion Matrix
  +---0----1----2----3----4----5----6----7----8----9--
 0|  43    1    0    1    2    1    2    2    2    1
 1|   0   43    0    0    2    6    0    0    0    4
 2|   1    0   47    2    2    0    2    1    0    0
 3|   1    0    1   50    0    0    1    2    0    0
 4|   0    1    0    0   54    0    0    0    0    0
 5|   0    3    0    0    0   47    0    1    0    4
 6|   0    0    1    0    0    0   51    0    3    0
 7|   3    0    1    0    1    0    0   49    1    0
 8|   0    0    0    1    0    0    3    0   51    0
 9|   1    3    0    0    0    4    0    0    0   47

 E_MS_DB_TM: nTrials= 550  nErr= 68   Prob[err]= 0.124
```

Figure 9. The result of ASTFT+CCs

Table.3 shows the probability of error that varies with different feature sets and their combinations.

Table 3. English and Chinese Prob[err] of different feature sets

	CCs	STFT	STFT+CCs	ASTFT	ASTFT+CCs
English PE	48.5%	20.5%	15.6%	13.6%	12.4%
Chinese PE	40.5%	27.8%	20.4%	19.5%	12.9%

After binarization, ASTFT+CCs gives the smallest Prob[err] while CCs gives the largest. English and Chinese Prob[err] of STFT+CCs are both less than that of single CCs and single STFT, which is true for ASTFT+CCs as well.

5.2 Results of neural network

When number of training epochs increases, the probability of error decreases first and increases again after epochs achieve a large number.Fig.22-24 show how two variables, PE for probability of error and PEV for probability of error of verification, change with number of epochs increasing. PEV means the probability of error that after using some spoken digits to train a neural network, taking the same spoken digits as input data. PEV should be close to 0 when the neural network is trained well, but not vice versa.

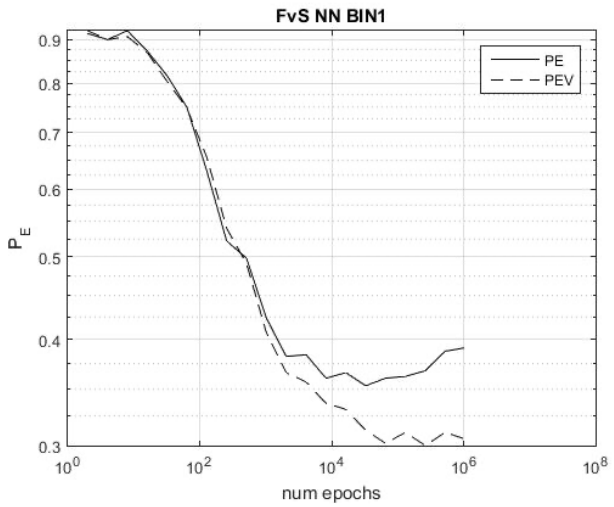
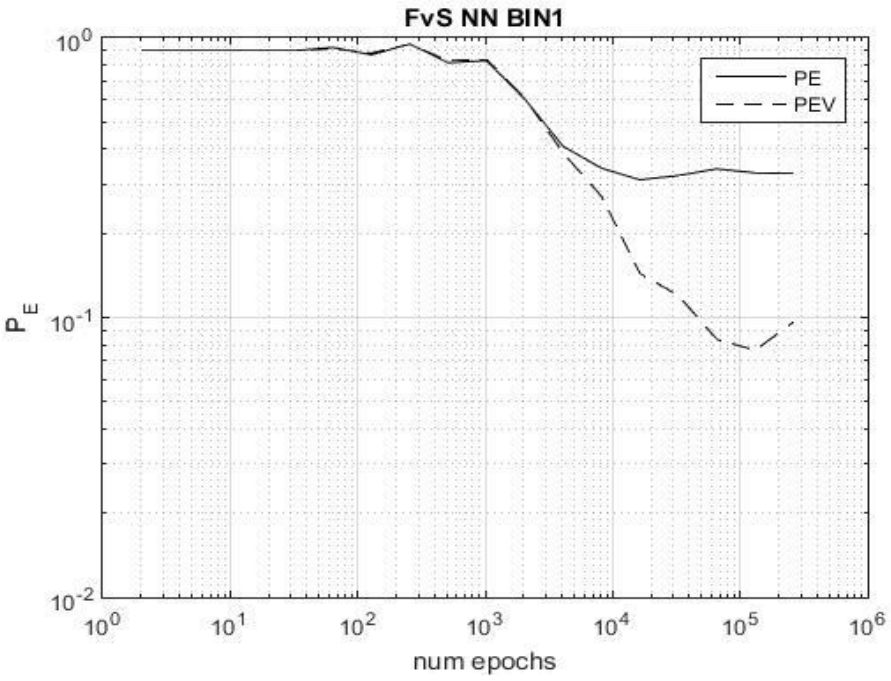


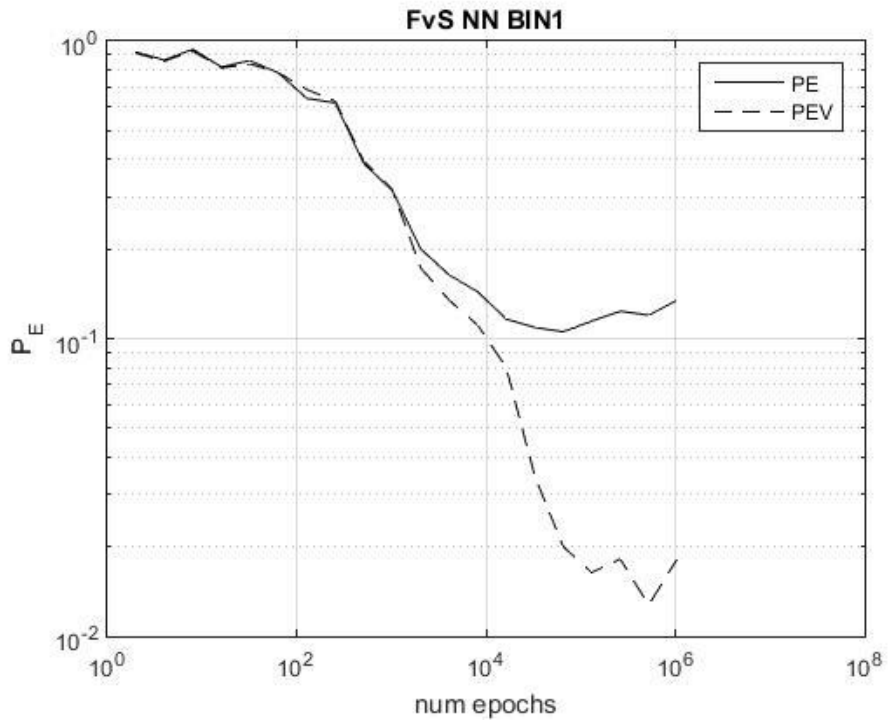
Figure 10. PE and PEV while the input vectors are raw spoken digits signal with one hidden layer

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

According to Fig.10, overtraining does occur in the training process because when the number of epochs is about 212, the minimum value of PEV is 0.3 which is much larger than 0. However, either the increasing number of hidden layer (from 1 to 2) or replacing the raw spoken digits, which was the input vectors of the neural network, with ASTFT+CCs, can overcome overtraining due to the obvious decrease of PEV. Fig.11 shows the results of the two circumstances. Fig.12 gives the minimum probability of error in Fig.11(b), which is slightly larger than that of template matching. The low sample size of training data is one possible reason.



(a)



(b)

Figure 11. (a) PE and PEV while the input vectors are raw spoken digits signal with two hidden layers. (b) PE and PEV while the input vectors are ASTFT+CCs with two hidden layers

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

E_MS_DB_NN2H Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	44	0	6	2	1	0	0	2	0	0
1	0	48	0	1	1	3	0	0	0	2
2	1	0	51	1	0	0	0	1	1	0
3	5	0	0	45	0	1	0	1	2	1
4	0	1	1	0	52	1	0	0	0	0
5	0	1	0	1	0	49	0	1	0	3
6	0	0	0	0	0	0	49	0	6	0
7	0	0	1	2	0	0	2	50	0	0
8	0	0	2	0	0	0	5	0	48	0
9	0	2	1	1	0	6	0	0	0	45

E_MS_DB_TM: nTrials= 550 nErr= 69 Prob[err]= 0.1255

Figure 12. Probability of error while training with ASTFT+CCs with two hidden layers

5.3 Result of multiple threshold and related method

Only using multiple thresholds surely can to some extent improve the classification result; however, such improvement is limited even when keeping on increasing the number of thresholds used. (Note: the data----probability of errors----is based on the best combination of thresholds randomly formed after many times) This phenomenon is showed by Figure. 13

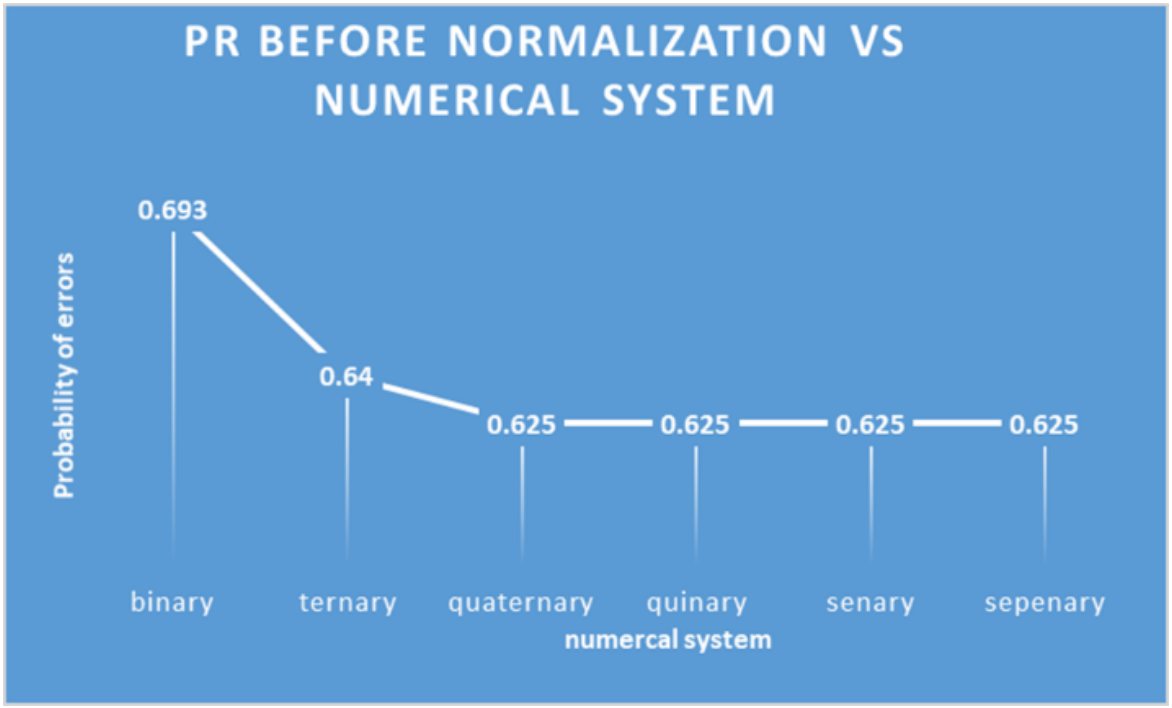


Figure 13. PR before normalization versus numerical system

While applying multiple thresholds, normalization, which would retrain the range of CCs back to 0-1, is also used. Thus, the thresholds we set would be more general for very sound. As Figure.14 shows, normalization is useful in reducing percentage of errors. It also shows that no matter applying normalization or not, the improvement led only by ad more thresholds is not useful when there are already 4 thresholds put in use.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041

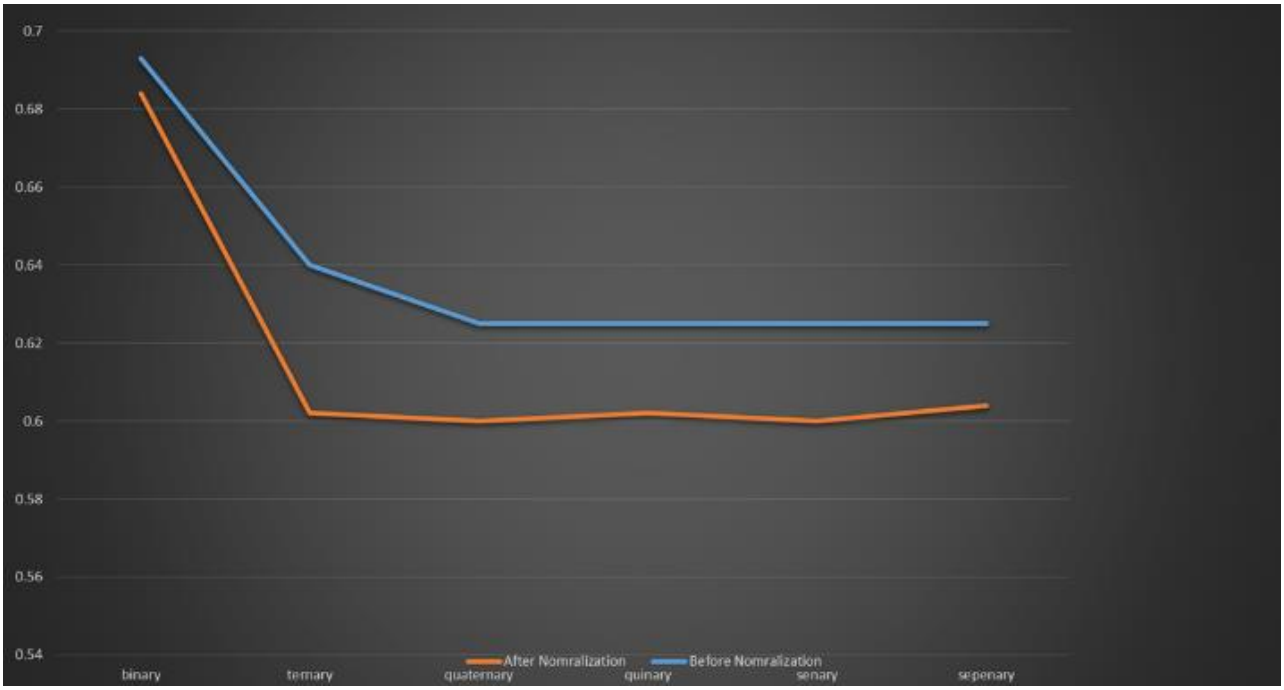


Figure 14. Normalization result

Although more thresholds are not useful anymore, making thresholds more precise would be a possible way to reduce the probability of errors (PE). As the thresholds become more preicse, more precise difference could be indentified. By comparing Fig. 15 and Fig. 16, it is clear that more decimal places added to the thresholds would be a good way to reduce PE.

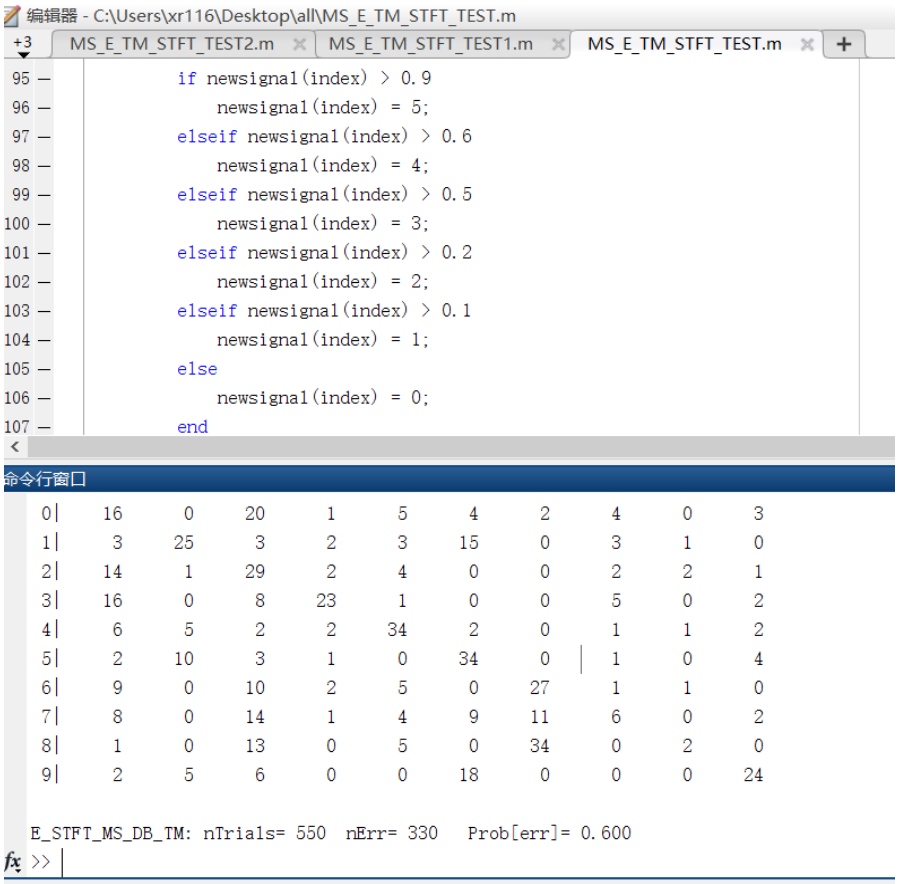
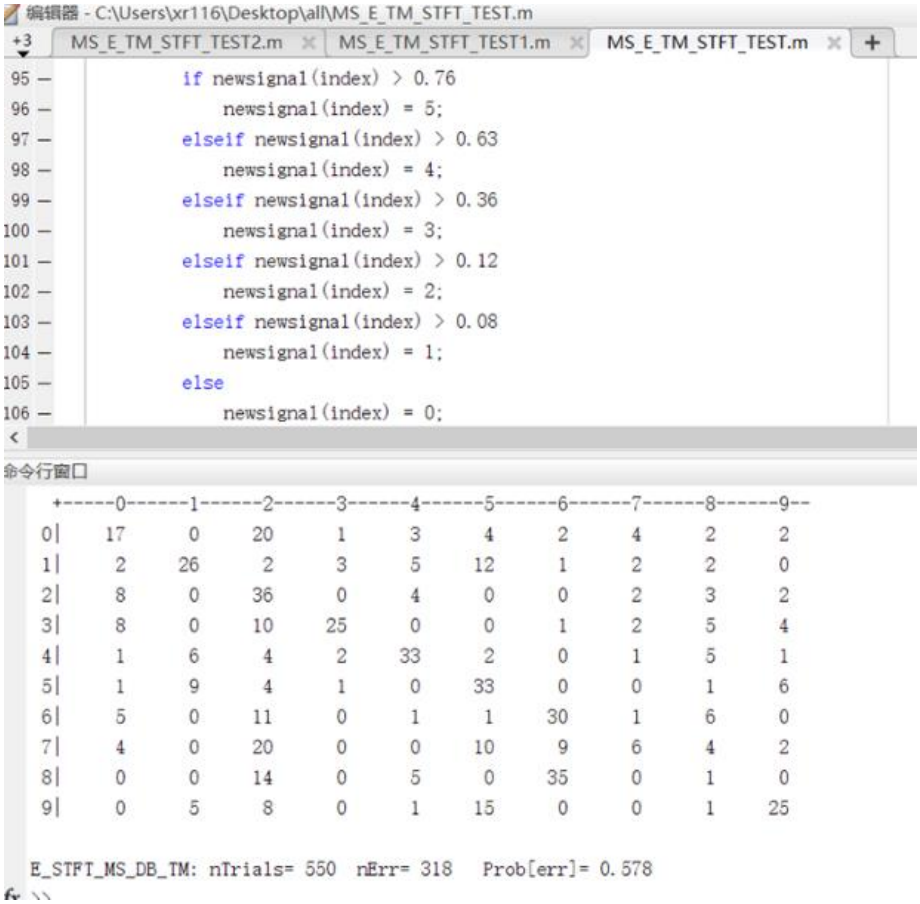


Figure 15. The probility of error when the thresholds are to the tenths decimal

ISSN: 2414-1895 DOI: 10.6919/ICJE.202101_7(1).0041



	0	1	2	3	4	5	6	7	8	9
0	17	0	20	1	3	4	2	4	2	2
1	2	26	2	3	5	12	1	2	2	0
2	8	0	36	0	4	0	0	2	3	2
3	8	0	10	25	0	0	1	2	5	4
4	1	6	4	2	33	2	0	1	5	1
5	1	9	4	1	0	33	0	0	1	6
6	5	0	11	0	1	1	30	1	6	0
7	4	0	20	0	0	10	9	6	4	2
8	0	0	14	0	5	0	35	0	1	0
9	0	5	8	0	1	15	0	0	1	25

Figure 16. The probility of error when the thresholds are to the hundreds decimal

6. Conclusion

This paper firstly presents ASTFT, a new feature set of spoken digits based on short-term Fourier transform. After experiments, the combination of ASTFT and CCs give the minimum using template matching after binarization. When artificial neural network is applied to decrease probability of error, overtraining happens. By training ASTFT+CCs with 2 hidden layers, the overtraining can be reduced whereas probability of error slightly increased comparing to that of template matching. Further investigation may aim to lower the probability of error of this newly trained neural network. Moreover, the paper introduces a new way to improve classification----multiple thresholds and manifests several factors that can further reduce the probability of errors.

References

[1] Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. J. of Computing. 2(3), 138–143 (2010).

[2] A. V. Oppenheim, “Speech spectrograms using the fast Fourier transform,” IEEE spectrum, vol. 8, no. 7, pp. 57–62, 1970.

[3] Hanchate, Dinesh & Nalawade, Mohini & Pawar, Manoj & Pophale, Vijay & Maurya, Prabhat. (2010). Vocal digit recognition using Artificial Neural Network. 6. V6-88 . 10.1109/ICCET.2010.5486314.

[4] L.R. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, 1993.