Speech Recognition and Optimization Using Linear Classification Artificial Neural Network

Jingbo Cui[1], Ting Liu[2*], Xinkai Hao[3*]

1 College of Information and Computer Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
2 College of electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
3 College of marine science and technology, Northwestern Polytechnical University, Xi'an 710072, China
**These authors contributed equally to this work and should be considered co-first authors.*

ABSTRACT. *This research studies the speech recognition process, and divides the speech recognition of linear system into four steps – speech acquisition, training, classification and results. For each part, its optimization is given. First, the effects of different feature sets of the same speech on classification results were tested. Then optimal parameter values of the neural network are found. Second, test the effect of different speech signal processing methods on speech recognition results. Present an analysis that shows whether STFT and ASTFT processing methods are effective in reducing error rate. Modify a neural network with four outputs to classify more digits. Third, the training step was modified from 10 outputs to 4 outputs (decimal to binary) and nCCs were transferred to binary for optimizing.*

KEYWORDS: *Neural network, Liner classification, Mscc, Stft*

1. Introduction

1.1 Background of Research

Speech is the acoustic expression of language. It is not only the most natural, effective and convenient way for humans to communicate but also an important tool for communication be- tween humans and machines. Visual speech recognition models traditionally consist of two stages, feature extraction and classification [1]. Through the analysis and processing of speech signals, feature extraction can remove redundant information which is not important to speech recognition and obtain the important in- formation that affects speech recognition. An artificial neural network (ANN) is a function that can be adjusted by training to classify objects from their feature vectors. This system has the characteristics of training, high parallelism, fast decision-making, and fault tolerance, which is suitable for speech signal processing. The artificial neural network method has a strong ability of self-organization and self-learning. When it is used in speech recognition, it has a strong ability of complex boundary resolution and robustness to incomplete information.

Artificial neural network is one of the most effective approaches for speech recognition thanks to its numerous architectures and learning algorithms [2]. Previous studies mainly focus on improving the model of artificial neural networks and often make great changes in structure to enhance the robustness of the neural network.

1.2 Problem Description

Intention of this research is to reduce the error rate of linear speech classification based on an artificial neural network. For a certain neural network, details optimizing the system need to be focused on. How to reduce the probability of error of this network will be the main problem in this research. Research will be carried out on three aspects.

First, the effect of different feature sets of a same speech signal database will be focused on. Next, the influence of different parameters on the result will be studied. The recognition results of different mode of artificial neural net- works will be looked at. After a series of tests and results analysis, conclusion is given, as well as an effective method to reduce the error rate of speech recognition.

Based on the original neural network with cepstrum coefficient (CCs) training, some optimization was made. Attention is transferred from current neural network with 10 outputs to a new neural network with 4 outputs (decimal to binary). This part mainly focused on CCs as obtained signal of speech for training.

Moreover, the existing research shows that prevalent ANNs have the same number of input and output, which is low-efficient. For instance, the NN that is used here has ten inputs and ten outputs. It can classify English or Chinese digits from 0-9. If more digits need to be classified, the neural network will have too many sub-layers with an increase in processing time. Thus, how to improve the efficient of a neural network in training and classification processes will be the problem in the last part.

2. Approach

The classification process of a linear neural network is divided into four steps, shown in Fig1.

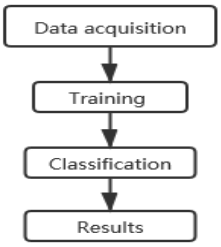


Fig.1 Procedure of Speech Recognition

2.1 Data Acquisition

Data acquisition is the procedure to collect the feature sets of speech signals. In this research, there are 3 different feature sets: CC with 30 feature vectors, spectrograms, using adaptive short-term Fourier transform with 100 feature vectors and the third one is the combination of these two sets above. The effects of these three different feature sets on the processing results will be compared in later work.

To overcome the limitation of Fourier Transform (FT) for analyzing nonstationary signals, and to provide excellent time and frequency localization of signal, Short-Time Fourier Transform (STFT) is used for better resolution in time and frequency domain than Fourier Transform.

STFT for fixed window length acts as a band-pass filter with fixed bandwidth through- out the time-frequency plane [3]. By hamming windowed, 10 time sections of the digit speech waveform are extracted.

2.2 Training

2.2.1 Parameters

For the training part, the influence of learning rate and the number of epochs on error rate are investigated.

First, the CCs are calculated, forms a MULTI-SPEAKER database of either English or Chinese spoken digit signals. Transform each row in speech database (DB) into a CC DB, and then form training DB and testing DB by adding layers in the way that odd-numbered layers into training database and even-numbered layers into Test DB. Then, ASTFT is conducted and form another feature set. Finally, cascade these two feature sets, to form a new combination DB, with 130 feature vectors in total.

A certain epoch was set, and the learning rate was gradually increased in each cycle for training, verification, and testing to obtain the curve between eta and error rate in Fig 2. Then keep eta unchanged, and gradually increase the number of epochs, to find the relation- ship between numRep and probability of error. Once the optimal eta and number of epochs are found, these certain values are used to train the network.

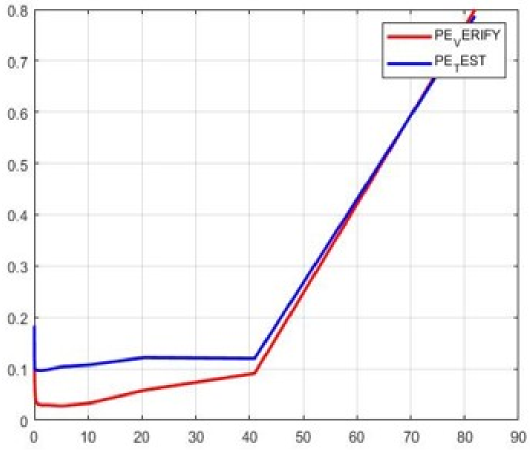


Fig.2 Curve of Eta and Pr[Err] for and Cc in Eng

2.2.2 Training Mode

The neural network has three stages, train, verify and test. First of all, original three stages were modified to let train stage has 4 outputs (0 or 1 for each), and let verify and test stages accept 4 outputs and transferred to decimal. For example, 0010 stands for digit 2 and 1001 stands for digit 9. It requires the four neural networks to set different target values. Second, the best threshold to transfer 4 outputs into one decimal in the verity threshold was observed. Third, in the train stage, CCs were transferred to binary to optimize the program. Fig 3 shows the results after transferring CCs to binary.

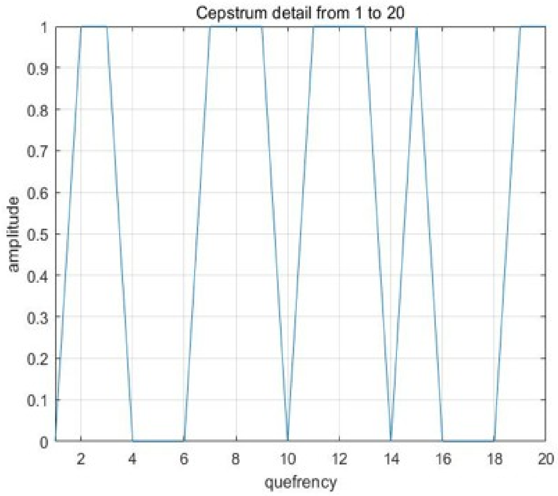


Fig.3 Transfer Ccs to Binary

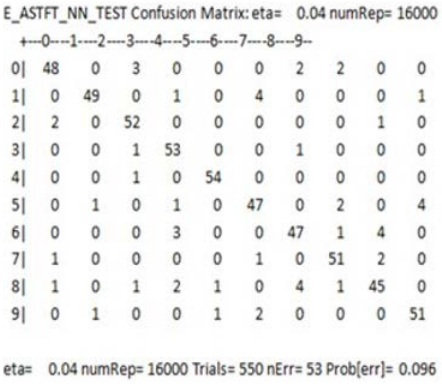
In addition, the diversity of English and Chinese were compared. Then train the NN with the database processed by Fourier transform and ASTFT, compare the difference between the two results.

2.3 Classification

Based on four-output neural network, more digit can be classified because 4-bit binary number can represent 16 digits in total. It would be more efficient for the whole system to classify the English and Chinese digits from 0 to 15 with less sub-neural network. The number of inputs is 16, using the speech from a single speaker. To see whether the result of the four-output neural network is better.

3. Results

3.1 Different Feature Sets



```
E_ASTFT_NN_TEST Confusion Matrix: eta=    0.04 numRep= 16000
  +---0---1----2----3----4----5----6----7----8----9--
 0|  48    0    3    0    0    0    2    2    0    0
 1|   0   49    0    1    0    4    0    0    0    1
 2|   2    0   52    0    0    0    0    0    1    0
 3|   0    0    1   53    0    0    1    0    0    0
 4|   0    0    1    0   54    0    0    0    0    0
 5|   0    1    0    1    0   47    0    2    0    4
 6|   0    0    0    3    0    0   47    1    4    0
 7|   1    0    0    0    0    1    0   51    2    0
 8|   1    0    1    2    1    0    4    1   45    0
 9|   0    1    0    0    1    2    0    0    0   51

eta=    0.04 numRep= 16000 Trials= 550 nErr= 53 Prob[err]= 0.096
```

Fig.4 Confusion Matrix for Astft Test

Different feature sets are tested respectively and their CM matrices are provided (Fig.4). A bar chart is drawn for easy observation, shown in Fig.5. The CC (Blue bar) set has the largest probability of error. CC is not sensitive to the location of the speech signal within the one-second acquisition window. However, it has no processing of original speech, which can be easily affected by low power noise

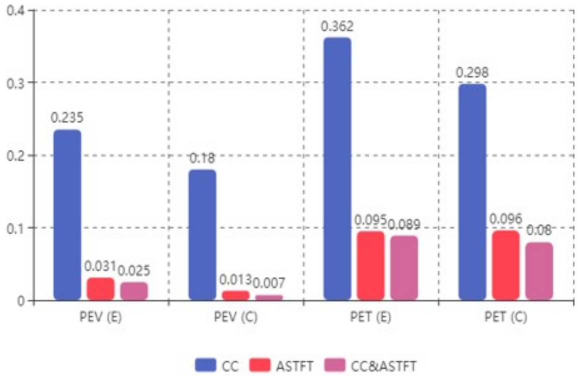


Fig.5 Results for Test and Verification

Note: PEV (E) = Probability of Error for Verify (English) PET (C) = Probability of Error for TEST (Chinese)

Adaptive short-term Fourier transform is the red bar in the figure, it first applies a threshold to the start and end of the speech waveform to determine the speech size, which may be different for each speech signal. ASTFT removes the small noise spike due to hitting the key be- fore recording, and it contains almost no zero values, which makes the features extremely efficient. That is the reason why ASTFT reduce the error significantly, for example, from 0.362 to 0.095. Detailed data is shown in Table 1.

Table 1 Data for Probability of Error

	CC	ASTFT	CC&ASTFT
PEV (E)	0.235	0.031	0.025
PEV (C)	0.18	0.013	0.007
PET (E)	0.362	0.095	0.089

PET (C)	0.298	0.096	0.08

The combination of two feature sets, the purple bar, has a further improvement than ASTFT. Because they can complement each other in features. For example, if a feature cannot be distinguished by ASTFT but can be identified by CC, the error caused by this feature will disappear. Moreover, the classification accuracy of the Chinese is generally higher than that of English. This may because most of us are native Chinese speakers and English pronunciation may be biased.

3.2 Training

3.2.1 Parameters

Part 1: Number of epoch (numRep)

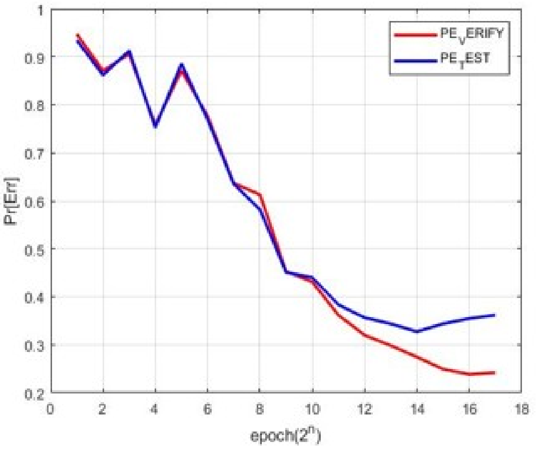


Fig.6 Curve of Epoch and Pr[Err] for Ccs in English

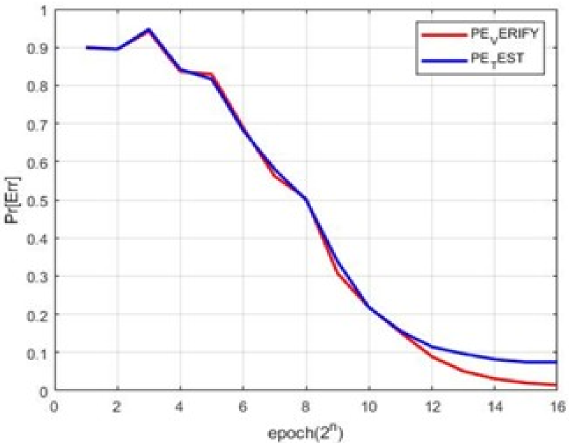


Fig.7 Curve of Epoch and Pr[Err] for Astft in English

Linear ANN can also be over-trained, shown in Fig 6 and 7 When there are too many epochs, the actual error rate will in- crease. With the increase of training times, hyperplane catches more training set value. The plane is moving around, trying to reduce the cost to include more points in the training set. However, not all points are useful and it may also include some wrong points. Thus, the probability of error will increase if there are too many epochs.

The number of epochs for 3 feature sets are tested separately, both for English and Chinese. The running time is proportional to the number of epochs, which means too many epochs will also cost a very long time. Meanwhile, too few training epochs will lead to insufficient training, which will cause a high error rate. Thus, data in Table 2 shows the optimal value ranges for each feature sets that provide sufficient training and acceptable running time.

Table 2 Optimal Epoch Range for Different Feature Sets

nRep	English	Chinese
CC	12000~16000	12000~16000
ASTFT	10000~14000	8000~12000
CC&ASTFT	8000~12000	12000~16000

Part 2: Learning rate (eta)

The process of searching for the best learning rate, the blue line is the error rate in the test. The eta that minimizes the PET (probability of error for the test), which is the lowest point shown in Fig 8, is chosen. Learning rates for different features are shown in Table 3. It shows that different database requires different eta value. Different learning rates lead to different error rates, which range from 0.9 to 0.1. This is an extremely vital factor. Thus, it is essential to test eta before training.

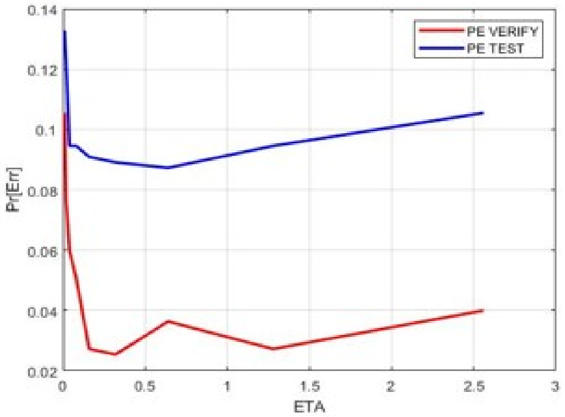


Fig. 8 Curve of Eta and Pr[Err] for Astft and Cc in Eng

Table 3 Optimal Eta for Different Feature Sets

eta	English	Chinese
CC	0.04	0.04
ASTFT	0.16	0.16
CC&ASTFT	0.64	0.64

3.2.2 Classification Mode

For the consistence of experiment, all train used eta= 10.00, Trials= 550, and epoch=10000. Before modifying the classification mode. In the test stage, the probability for English is 0.378, and 0.338 for Chinese. Results for 4 outputs program for CC is 0.599 error rate for Chinese and 0.617 for English. As for 4 outputs program with binary CCs, error rate for English is 0.690, while 0.713 for Chinese.

◊Best threshold transfer 4 outputs to decimal

Fig 9 shows the best threshold of English is 0.7, Fig 10 shows the best threshold of Chinese is 0.55. However, based on several observations of testing, the best threshold will change little along with different train stage.

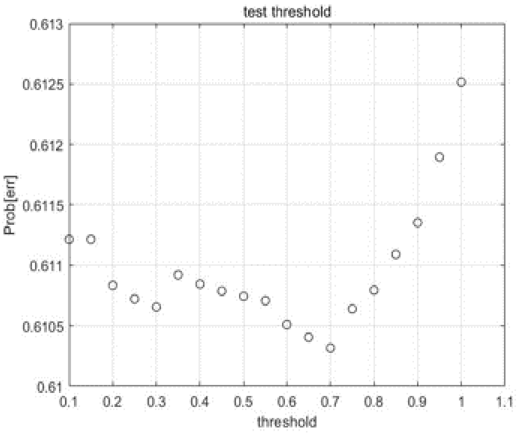


Fig.9 English Threshold

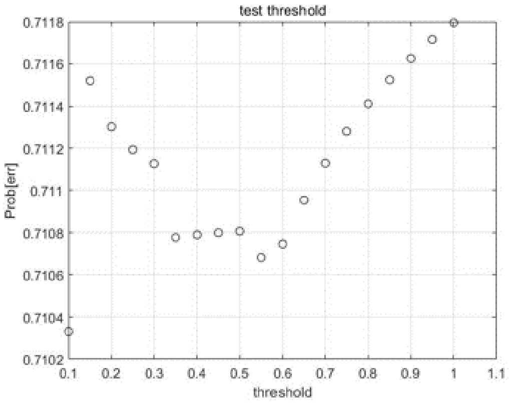
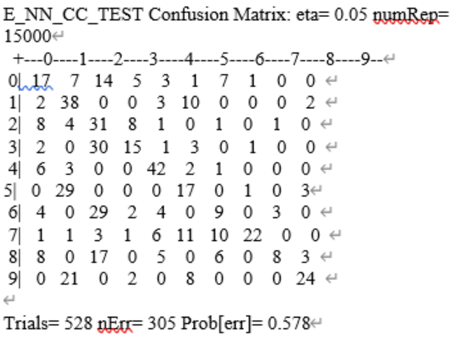


Fig.10 Chinese Threshold

◊Outputs ANN with ASTFT.

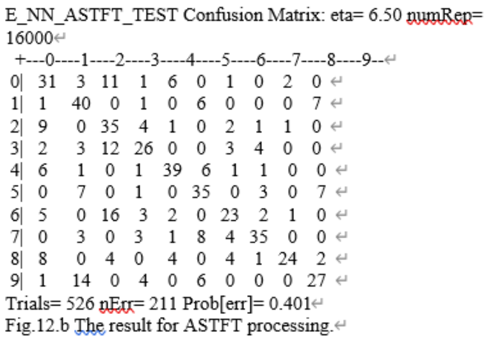
In this section, different English digits test results of the four-output ANN recognition are presented for different ways to process the database. The performance of the ANN is investigated for FT and ASTFT. The Fig.11 and Fig.12 shows the results. Both eta and numRep are optimal values found by the process described below.



```
E_NN_CC_TEST Confusion Matrix: eta= 0.05 numRep=
15000
 +---0----1----2----3----4----5----6----7----8----9--
 0| 17   7  14   5   3   1   7   1   0   0
 1|  2  38   0   0   3  10   0   0   0   2
 2|  8   4  31   8   1   0   1   0   1   0
 3|  2   0  30  15   1   3   0   1   0   0
 4|  6   3   0   0  42   2   1   0   0   0
5|  0  29   0   0   0  17   0   1   0   3
 6|  4   0  29   2   4   0   9   0   3   0
 7|  1   1   3   1   6  11  10  22   0   0
 8|  8   0  17   0   5   0   6   0   8   3
 9|  0  21   0   2   0   8   0   0   0  24

Trials= 528 nErr= 305 Prob[err]= 0.578
```

Fig.11 the Classification Result for Ft Processing

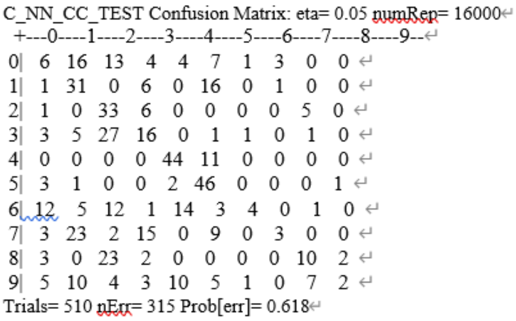


```
E_NN_ASTFT_TEST Confusion Matrix: eta= 6.50 numRep=
16000
 +---0----1----2----3----4----5----6----7----8----9--
0|  31   3  11   1   6   0   1   0   2   0
1|  1   40   0   1   0   6   0   0   0   7
2|  9    0  35   4   1   0   2   1   1   0
3|  2    3  12  26   0   0   3   4   0   0
4|  6    1   0   1  39   6   1   1   0   0
5|  0    7   0   1   0  35   0   3   0   7
6|  5    0  16   3   2   0  23   2   1   0
7|  0    3   0   3   1   8   4  35   0   0
8|  8    0   4   0   4   0   4   1  24   2
9|  1   14   0   4   0   6   0   0   0  27
Trials= 526 nErr= 211 Prob[err]= 0.401
Fig.12.b The result for ASTFT processing.
```

Fig.12 the Classification Result for Astft Processing

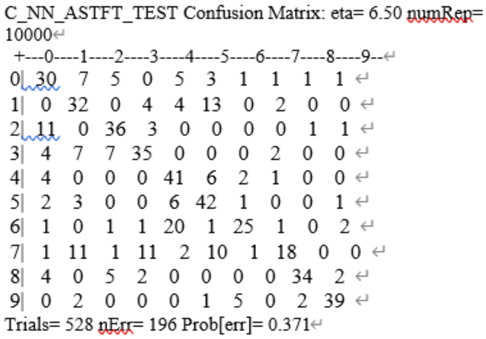
From the confusion matrices above, the Prob[err] of ASTFT processing is lower than the FT. The Prob[err] of ASTFT is 0.401. The number of Trials is not 550, because the program got rid of the wrong data. If the recognition result is not between 0000 and 1000 (0 to 9), we consider the result error and remove the result of this error. ASTFT can improve the above deficiencies, and the window is adaptive. The formal part of the research found that ASTFT can distinguish the feature of signals more clearly and significantly reduce the error rate.

Then classifications of the Chinese digits are also conducted. The Fig 13 and Fig.14 shows the results.



```
C_NN_CC_TEST Confusion Matrix: eta= 0.05 numRep= 16000
 +---0----1----2----3----4----5----6----7----8----9--
0|   6  16  13   4   4   7   1   3   0   0
1|   1  31   0   6   0  16   0   1   0   0
2|   1   0  33   6   0   0   0   0   5   0
3|   3   5  27  16   0   1   1   0   1   0
4|   0   0   0   0  44  11   0   0   0   0
5|   3   1   0   0   2  46   0   0   0   1
6|  12   5  12   1  14   3   4   0   1   0
7|   3  23   2  15   0   9   0   3   0   0
8|   3   0  23   2   0   0   0   0  10   2
9|   5  10   4   3  10   5   1   0   7   2
Trials= 510 nErr= 315 Prob[err]= 0.618
```

Fig.13the Result for Cc Processing for Chinese Digits



```
C_NN_ASTFT_TEST Confusion Matrix: eta= 6.50 numRep=
10000
 +---0----1----2----3----4----5----6----7----8----9--
0|  30   7   5   0   5   3   1   1   1   1
1|   0  32   0   4   4  13   0   2   0   0
2|  11   0  36   3   0   0   0   0   1   1
3|   4   7   7  35   0   0   0   2   0   0
4|   4   0   0   0  41   6   2   1   0   0
5|   2   3   0   0   6  42   1   0   0   1
6|   1   0   1   1  20   1  25   1   0   2
7|   1  11   1  11   2  10   1  18   0   0
8|   4   0   5   2   0   0   0   0  34   2
9|   0   2   0   0   0   1   5   0   2  39
Trials= 528 nErr= 196 Prob[err]= 0.371
```

Fig.14 the Result for Astft Processing for Chinese Digits

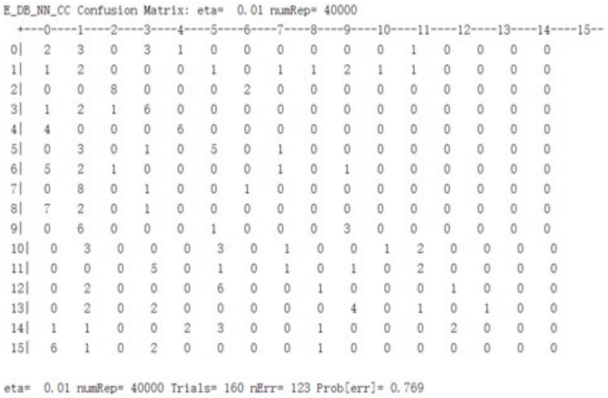
ASTFT processing also plays an important role in Chinese digit speech recognition. It can reduce about half of the error rate, from 0.618 to 0.371.

During the training process, the time of the binary neural network training is much shorter than the time of the decimal output for the same number of the epoch. The binary network does not need too many training times, with a small number of training epochs that can reach to the highest accuracy. A problem is that the NN with four-outputs works not as well as the NN with ten-output.

3.3 Classification

◊The result of English digits 0-15 classification

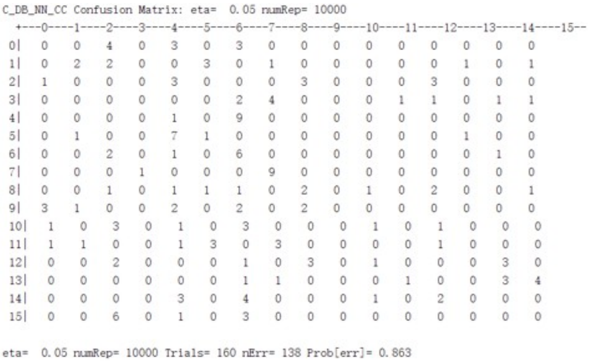
The difference between the NN with four-outputs and NN with ten-outputs is that binary number is used instead of the decimal number. Through this step, the range of identifying numbers is expanded from 0-9 to 0-15. This is a symbol of increasing the number of outputs in a wider application range. Here, a speech database for English and Chinese digits from 0-15. The Fig 15 and Fig 16 shows the result of the classification.



```
E_DB_NN_CC Confusion Matrix: eta=  0.01 numRep= 40000
 +---0----1----2----3----4----5----6----7----8----9----10----11----12----13----14----15--
 0|   2    3    0    3    1    0    0    0    0    0    0    1    0    0    0    0
 1|   1    2    0    0    0    1    0    1    1    2    1    1    0    0    0    0
 2|   0    0    8    0    0    0    2    0    0    0    0    0    0    0    0    0
 3|   1    2    1    6    0    0    0    0    0    0    0    0    0    0    0    0
 4|   4    0    0    0    6    0    0    0    0    0    0    0    0    0    0    0
 5|   0    3    0    1    0    5    0    1    0    0    0    0    0    0    0    0
 6|   5    2    1    0    0    0    0    1    0    1    0    0    0    0    0    0
 7|   0    8    0    1    0    0    1    0    0    0    0    0    0    0    0    0
 8|   7    2    0    1    0    0    0    0    0    0    0    0    0    0    0    0
 9|   0    6    0    0    0    1    0    0    0    3    0    0    0    0    0    0
 10|   0    3    0    0    0    3    0    1    0    0    1    2    0    0    0    0
 11|   0    0    0    5    0    1    0    1    0    1    0    2    0    0    0    0
 12|   0    2    0    0    0    6    0    0    1    0    0    0    1    0    0    0
 13|   0    2    0    2    0    0    0    0    0    4    0    1    0    1    0    0
 14|   1    1    0    0    2    3    0    0    1    0    0    0    2    0    0    0
 15|   6    1    0    2    0    0    0    0    1    0    0    0    0    0    0    0

eta=  0.01 numRep= 40000 Trials= 160 nErr= 123 Prob[err]= 0.769
```

Fig.15 the Result of English Digits 0-15 Classification



```
C_DB_NN_CC Confusion Matrix: eta=  0.05 numRep= 10000
 +---0----1----2----3----4----5----6----7----8----9----10----11----12----13----14----15--
 0|   0    0    4    0    3    0    3    0    0    0    0    0    0    0    0    0
 1|   0    2    2    0    0    3    0    1    0    0    0    0    0    1    0    1
 2|   1    0    0    0    3    0    0    0    3    0    0    0    3    0    0    0
 3|   0    0    0    0    0    0    2    4    0    0    0    1    1    0    1    1
 4|   0    0    0    0    1    0    9    0    0    0    0    0    0    0    0    0
 5|   0    1    0    0    7    1    0    0    0    0    0    0    0    1    0    0
 6|   0    0    2    0    1    0    6    0    0    0    0    0    0    0    1    0
 7|   0    0    0    1    0    0    0    9    0    0    0    0    0    0    0    0
 8|   0    0    1    0    1    1    1    0    2    0    1    0    2    0    0    1
 9|   3    1    0    0    2    0    2    0    2    0    0    0    0    0    0    0
 10|   1    0    3    0    1    0    3    0    0    0    1    0    1    0    0    0
 11|   1    1    0    0    1    3    0    3    0    0    0    0    1    0    0    0
 12|   0    0    2    0    0    0    1    0    3    0    1    0    0    0    3    0
 13|   0    0    0    0    0    0    1    1    0    0    0    1    0    0    3    4
 14|   0    0    0    0    3    0    4    0    0    0    1    0    2    0    0    0
 15|   0    0    6    0    1    0    3    0    0    0    0    0    0    0    0    0

eta=  0.05 numRep= 10000 Trials= 160 nErr= 138 Prob[err]= 0.863
```

Fig.16 the Result of Chinese Digits 0-15 Classification

When binary NN is used to classify more digits, the result of classification is bad. The Prob[err] even reaches 0.863, which means the network requires a deeper training. There is a lot of space for improvement.

4. Conclusion

For data acquisition, the combination of CC and ASTFT is a better choice. With proper processing of the feature sets of speech signals, the accuracy of a system can be greatly improved. The key is to find and get more informative feature vectors. This study only tested small data sets, and it can be inferred that when the data set is large, increasing the number of feature vectors may result in huge consumption. Thus, it is worth trying to reduce running consumption by reducing features with lower values and less impact, which is also called ‘data pooling’. This is especially urgent for companies using large data sets.

In the training step, parameters (Both eta and numRep) can significantly affect the accuracy of classification results, in other words, the probability of error, so it is necessary to find the optimal parameters values of a linear neural network, or maybe other similar systems. Though the best threshold to convert 4 outputs to one decimal was observed, transfer 10 outputs to 4 outputs (decimal to binary) decreases the accuracy of classification, and the results of transfer CCs to binary based on 4 outputs are worse. In the future study, different thresholds can be set based on the CCs of different numbers that may optimize the program.

Even though the accuracy of classification was decrease after modifying 10 outputs to binary 4 outputs. However, it improves the speed and efficiency of the classification with fewer sub-networks and greatly reduces the computation time, which means it has more practical uses for the hardware. This method can be further developed in subsequent research.

References

[1] Maruf A.Dhali, Camilo Nathan Jansen, Jan Willem de Wit, Lambertb Schomaker (2020). Feature-extraction methods for historical manuscript dating based on writing style development. Pattern Recognition Letters, vol.131, no3, pp.2020.

[2] Nawel SOUISSI, Adnane CHERIF (2016). Speech Recognition System Based on Short-term Cepstral Parameters, Feature Reduction Method and Artificial Neural Networks. 2nd International Conference on Advanced Technologies for Signal and Image Processing ATSIP', no.3, pp.21-24.

[3] Anukul Anand, Manoj Kumar Mukul (2016). Comparison of STFT Based Direction of Arrival EstimationTechniques for Speech Signal. IEEE International Conference on Recent Trends in Electronics Information Communication Technology, no.5, pp.20-21.