INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

ToneNet: A CNN Model of Tone Classification of Mandarin Chinese

Qiang Gao, Shutao Sun, Yaping Yang*

School of Computer and Cyberspace Security, Communication University of China
Beijing, China

qiangao, stsun, yyp_berry@cuc.edu.cn

Abstract

In Mandarin Chinese, correct pronunciation is the key to convey word meaning correctly and the correct pronunciation is closely related to the tone of text. Therefore, tone classification is a critical part of speech evaluation system. Traditional tone classification is based on F0 and energy or MFCCs. But the extraction of these features is often subject to noise and other uncontrollable environmental factors. Thus, in order to reduce the influence of environment, we designed a CNN network named ToneNet which adopts mel-spectrogram as a feature and uses a customed convolutional neural network and multi-layer perceptron to classify Chinese syllables into one of the four tones. We trained and tested ToneNet on the Syllable Corpus of Standard Chinese Dataset (SCSC). The result shows that the best accuracy and f1-score of our method have reached 99.16% and 99.11% respectively. Besides, ToneNet has achieved 97.07% of accuracy and 96.83% of f1-score with the condition of gaussian noise.

Index Terms: ToneNet; tone classification; mel-spectrogram; Mandarin Chinese; convolutional neural networks

1. Introduction

Tone classification plays an important role in speech evaluation of Mandarin Chinese. Different tones of the same Chinese character usually have different meanings. Tone can increase the intelligibility of Mandarin Chinese sentences. [5] indicates that the accuracy of Mandarin Chinese sentence recognition will be greatly reduced when the tone information of the sentence is interfered by noise. Therefore, tone classification is not only helpful to reduce the error rate of speech recognition system, but also helpful to improve the evaluation ability of speech evaluation system. In addition, the tone of Chinese character also plays a crucial role in speech synthesis. The correct tone can make the synthesis of Mandarin Chinese more natural.

Mandarin Chinese has four different tones: (1) flat and high, (2) rising, (3) low and dipping, and (4) falling [1]. The four types of tone have their own unique contour. So it can be distinguished accurately by selecting appropriate feature. Traditional methods mainly use F0 and energy as features [2] [3]. Although F0 and energy of speech can reflect the prosodic feature of monosyllabic Mandarin Chinese, they are influenced easily due to the interference of environmental noise. Ryant et al [16] train a DNN model based on 40 MFCCs and it achieved 27.38% frame error rate (FER) and 15.62% segment error rate (SER). Besides, Chen et al [1] use MFCCs as the input feature and the accuracy has achieved 95.53%. The method of [1] also uses a CNN model. Although Chen et al [1] use denoising autoencoder (dAE) to pre-trained the features, it still not overcome the influence of the noise completely. In this paper, we propose a new network named ToneNet. ToneNet uses mel-spectrogram of speech as classification feature and customed convolution neural network as the extractor of feature. On the Syllable Corpus of Standard Chinese dataset (SCSC), our method has achieved 99.16% of accuracy and 99.11% of f1-score. Figure 1 shows the mel-spectrogram of the four tones and the typical contour of each tone is clearly visible. So, we use low frequency region of mel-spectrogram to reduce the influence of environmental noise effectively and achieve a good classification performance.

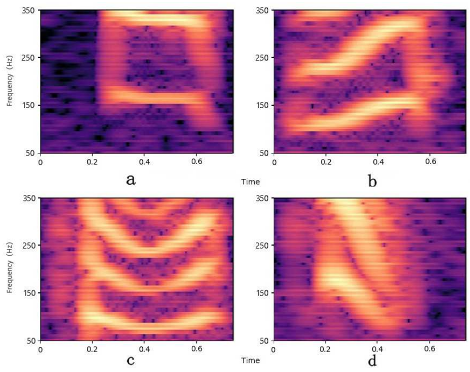

Figure 1: *Four tones mel-spectrogram of Chinese syllable. a: flat and high. b: rising. c: low and dipping. d: falling.*

The main contributions of this paper mainly focus on the following two aspects：

- Model architecture: An efficient 5-layer convolutional neural network is designed as a feature extractor and a 3-layer multi-layer perceptron is used as a classifier.
- Feature processing: Low frequency region of mel-spectrogram is used as the feature of tone classification. We save it as image for input of model.

2. Feature preprocess

Traditional feature F0 is used for tone classification is disturbed by environmental noise easily and it is unstable to extract it. At the same time, it is easy to cause gradient explosion or non-convergence when acoustic features are used for training directly in deep learning. Mel-spectrogram contains more raw information than f0. Therefore, this paper proposes to use mel-spectrogram for tone classification. Mel-spectrogram not only takes into account the auditory habits of the human's ear, but

[*] Corresponding author

http://dx.doi.org/10.21437/Interspeech.2019-1483

also reflects the energy feature when it is converted to a spectrogram. Tone classification has nothing to do with the content of speech. It does not focus on the meaning of speech, but only on the prosody of speech. From the Figure 2. a we can observe tone contour (prosody) in the range of low frequency. The range of mel-spectrogram's full frequency is [0, 8000] Hz in this paper. As frequency increases, we are unable to distinguish tone information increasingly. Thus, the main information of tone is concentrated in the low frequency area of black rectangular box on the right of Figure 2. b. It has little interference information in the black rectangular box because we only select the low frequency area of mel-spectrogram and the tone contour is the most obvious information in the region.

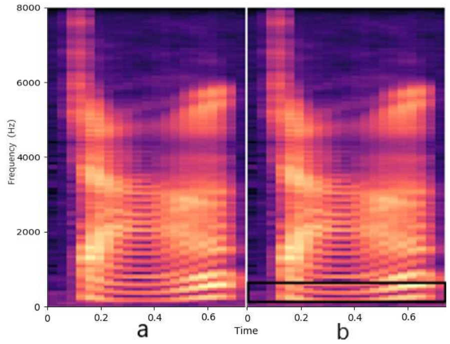

Figure 2: *Tone mel-spectrogram: low and dipping.*

Human's voice is multi-harmonic. The range of F0 is about 100 ~ 300 Hz when people speak. The F0 of male is lower, the F0 of female and child is higher. So the range of low frequency (the black rectangle) which we choose is [50, 350] Hz in this paper in order to cover the range of human tones' F0. At last, we save this area as image for input of model.

3. ToneNet architecture

Convolutional neural network (CNN) is a feed forward neural network, whose artificial neurons can respond to the surrounding units within a part of the coverage, and it performs image processing excellently [6]. CNN has been widely used in various fields, such as image classification, target detection, speech recognition. Many researchers have proposed a lot of variant networks based on CNN, such as AlexNet [7], VGGNET [8], RESNET [9]. These networks which are based on CNN not only deepen the depth of the convolutional neural network, but also more increased its ability of feature extraction and further extend the application of CNN.

MLP (Multilayer Perceptron) is also known as ANN (Artificial Neural Network). In addition to input and output layer, it may have multiple hidden layers in the middle. The simplest MLP consists of only one hidden layer, namely three-layer structure. We can see from Figure 3 that the multi-layer perceptron is composed of fully connected layers.

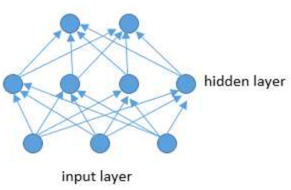

Figure 3: *Three-layer structure of the simplest MLP.*

In other words, the multi-layer perceptron's adjacent layers are fully connected with each other. MLP has been proved that it is a general function of approximation method, which can be used to fit complex functions or solve problems of classification.

This paper designs a network based on CNN and MLP to perform tone classification named ToneNet. Figure 4 shows that TonNet consists of 5 2D convolution layers and 3 fully connected layers. After each convolution layer, it is the BatchNormalization layer and the 2D MaxPooling layer. The convolution layer is responsible for the extraction of image features. The fully connected layer is used to fit the features that are extracted from the convolution layer and map them to the corresponding categories by using softmax.

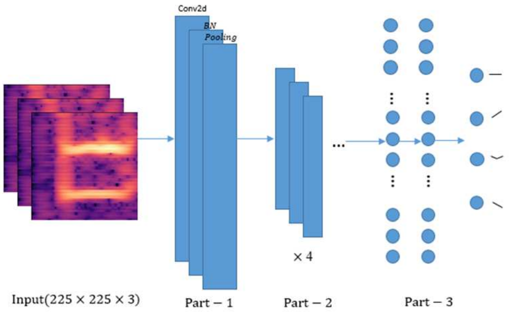

Figure 4: *The ToneNet structure: Three major modules: Part-1, Part-2 and Part-3. Part-1 includes Conv2d with a convolution kernel of 5 × 5, BatchNormalization (BN) and MaxPooling. Part-2 is similar to Part-1 but its convolution kernel is 3 × 3 and there are four of the same layers. Part-3 is a three-layer structure of the simplest MLP.*

ToneNet is mainly made up of three major modules. The detailed model structure is shown in Table 1. The input of ToneNet is an RGB image of mel-spectrogram.

The first module of ToneNet (part-1) consists of a convolution layer, a batch normalization layer and a maxpooling layer. The size of the convolution kernel in the first convolution layer of Part-1 is 5 × 5, the number of convolution kernel is 64 and the stride is 3. As the input layer, the first convolutional layer mainly plays the role of parameter dimension reduction and preliminary extraction of feature information.

The second layer is the batch normalization layer. Batch normalization layer was proposed by [10] in 2015. Its main function is to process the input data of each layer of neural network. Therefore, batch normalization layer not only speeds up the convergence of the network greatly, but also prevents overfitting in the model training of deep neural network effectively. The third layer is the maxpooling layer. The size of the filter for the maxpooling layer is 3 × 3 and the stride is 3. The size of output is 25 × 25 × 64 in the first module.

The second module of ToneNet (part-2) consists of four convolution layers, four batch normalization layers and four maxpooling layers. The convolutional layers of Part-2 are different from these of Part-1. The size of the convolution kernel of Part-2 is 3 × 3 with a stride of 1. The size of the filter of the maxpooling layer is 2 × 2 with a stride of 2. The size of the convolution kernel and pooling kernel of Part-2 is smaller than that of Part-1. Part-2 of ToneNet follows the design idea of VGGNET [8]. The authors of VGGNET used two convolution kernels with the size of 3 × 3 instead of one convolution kernel with the size of 5 × 5 in the shallow networks and the result showed that the feature extraction effect of multiple small convolution kernels was better than that of single large convolution kernel [8]. The size of output is 2 × 2 × 512 in Part-2.

Table 1: *The ToneNet architecture, the f is the size of convolution kernels and the s is stride.*

Name	**ToneNet**
Input	Image

Part-1	Conv2d(f=5 × 5 × 64, s=3)
	BatchNormalization
	MaxPooling2d(f=3 × 3, s=3)

Part-2	Conv2d(f=3 × 3× 128, s=1)
	BatchNormalization
	MaxPooling2d(f=2 × 2, s=2)
	Conv2d(f=3 × 3 × 256, s=1)
	BatchNormalization
	MaxPooling2d(f=2 × 2, s=2)
	Conv2d(f=3 × 3 × 256, s=1)
	BatchNormalization
	MaxPooling2d(f=2 × 2, s=2)
	Conv2d(f=3 × 3 × 512, s=1)
	BatchNormalization
	MaxPooling2d(f=2 × 2, s=2)

Flatten	Flatten
Part-3	FC-1024 BatchNormalization FC-1024 BatchNormalization FC-4 SoftMax

The flatten layer which follows Part-2 aims to connect convolution layer to fully connected layer. The third module of ToneNet (part-3) is a multi-layer perceptron that consists of three fully connected layers and two batch normalization layers. The first two fully connected layers of Part-3 have 1024 neurons, and the last fully connected layer has 4 neurons which correspond to the four categories of tone.

The activation function of ToneNet are ReLU (Rectified Linear Units). The definition of ReLU is that it makes the value less than 0 equal to 0, otherwise it leaves the value unchanged. This is a crude way to force some data to be 0 in formula 1.

$$f(x) = max(0, w^T x + b) \tag{1}$$

x is the output of the previous layer, W^T is the weight of the current layer，and *b* is the bias value. The practice indicates that it gives moderate sparsity for network and improves the training speed greatly [11]. The loss function of ToneNet is the categorical cross-entropy function, and the optimizer is the Stochastic Gradient Descent (SGD) with the momentum and nesterov. SGD algorithm can make better use of information, especially when information is redundant [12].

4. Experiment of ToneNet and analysis of result

In this experiment, the dataset we used is Syllable Corpus of Standard Chinese Dataset (SCSC). The creator is The Institute of Linguistics Chinese Academy of Social Sciences. Mandarin mono-syllable corpus is comprised by mono-syllable wave data, the list of mono-syllable and management software. The SCSC dataset consists of 1,275 monosyllabic Chinese characters, which are composed of 15 pronunciations of young men. So it has a total of 19,125 pronunciations. The audio format is mono and 16bit WAV and its sample rate is 16,000 Hz. The duration of each audio is about 0.5 ~ 1s. The SCSC dataset was divided into training set, validation set and test set according to the ratio of 8:1:1. Training set and validation set are used to train and adjust parameters. Finally, the test set is used to test the performance of the final model. The evaluation metrics are accuracy, precision, recall and f1-score. F1-score reflects the overall performance of precision and recall rate, as is shown in formula 2.

$$f1 = \frac{2 * precision * recall}{precision+recall} \tag{2}$$

Formula 2 shows that f1 integrates the results of precision and recall, and when f1 is higher, ToneNet is more effective.

From Figure 1 and Figure 2 we can see that low frequency seems to be more effective for tone classification. However, it is difficult to decide whether the region outside the black rectangular box should be discarded. So we have designed six group experiments to choose a more efficient mel-spectrogram for tone classification. The experiment settings are as follows:

- Frequency: The full frequency and low frequency (black rectangle box) are selected respectively, namely the range of low frequency is [50, 350] Hz and the range of full frequency is [0, 8000] Hz. The purpose is that we verify the influence of the region outside the black rectangle box on the tone classification.
- Size: (113, 113, 3), (225, 225, 3) and (449, 449, 3) are selected for analyzing the influence of picture size on the tone classification.

Based on the above conditions, six group experiments were performed. In the extraction stage of mel-spectrogram, in order to achieve a clearer mel-spectrogram under the low frequency range of [50, 350] Hz, 64 mel filters are used. The frame length is 2048 sampling points, and the frame shift is 16 sampling points. Mel-spectrogram is saved as a picture. The feature extraction tool is librosa which is a professional voice processing library [15]. In the model training stage, our basic learning rate was set to 0.001, mini-batch was set to 128, and epoch was set to 50. All the results are shown in Table 2. The six group experiments used the same model and parameters and the SCSC were used as the dataset. As is shown in Table 2, the result in low frequency and the size of (225, 225, 3) is the best. We can see that the test loss in the low frequency is lower than this in full frequency commonly.

Although the improvement of model which use the low-frequency region as feature is smaller, Low frequency is still more beneficial to the tone classification. The size of (225, 225, 3) and (449, 449, 3) with low frequency achieve a similar result. In order to save memory and improve training speed, this paper adopts mel-spectrogram with frequency range of [50, 350] Hz and image size of (225, 225, 3) as the input of the customed convolutional neural network model like Figure 1. The experiment result shows that mel-spectrogram in low frequency has a good feature expression ability in the tone classification. Whether ToneNet pays attention to the prosodic feature of Chinese pronunciation and learns it is one of the key point of our research. Therefore, we use Grad-cam [14] to visualize the concerns of ToneNet in tone classification. The last convolution layer of ToneNet is converted into a heat map and applied to the original image. As is shown in Figure 5.

Table 2: *Accuracy, F1-score and test loss of six group of experiments.*

Category	Accuracy （%）	F1-score （%）	Test Loss
low frequency with size of (113, 113, 3)	97.90	97.83	0.06207
low frequency with size of (225, 225, 3)	**99.16**	**99.11**	**0.05153**
low frequency with size of (449, 449, 3)	99.00	98.93	0.05312
full frequency with size of (113, 113, 3)	96.86	96.68	0.08151
full frequency with size of (225, 225, 3)	97.73	97.57	0.07082
full frequency with size of (449, 449, 3)	97.75	97.60	0.07011

There are 16 pictures in Figure 5. Each line has two pronunciations of the same tone. The first and third columns are the original images. The second and the fourth column are the images using the heat map. Figure 5 shows that ToneNet focused on the brightness area of the original image which reflects the contour of each tone. This attention mechanism is the same as our human being on vision.

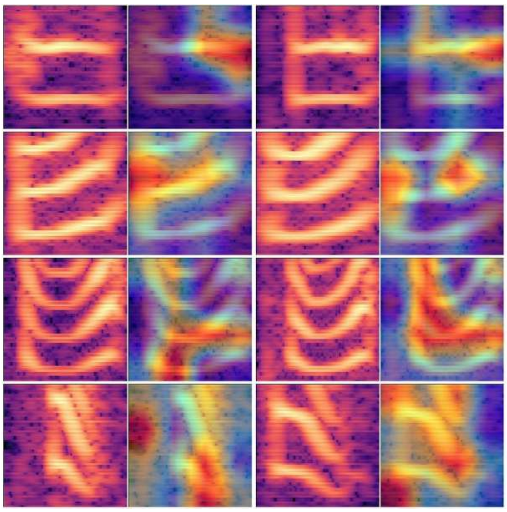

Figure 5: *The four tones and its heat maps. Each line has two pronunciations of the same tone.The first and third columns are the original images. The second and the fourth column are the images using the heat map.*

Table 3 shows the best experimental result of ToneNet and the result of the other networks. ToneNet has the best performance in all evaluation metrics. The accuracy of [13] only is 72.8%. The performance of [1] is closest to that of ToneNet, but it does not provide more representative evaluation metrics of the model, such as precision, recall and f1-score.

Table 3: *Experimental result of ToneNet and two related models. Acc is the accuracy, P is the precision, R is the recall and F1 is the f1-score.*

System	Acc	P	R	F1	Data
[13]	72.80	-	-	-	MCCC
[1]	95.53	-	-	-	MCCS
[1]	94.45	93.51	94.63	94.06	SCSC
ours	**99.16**	**99.08**	**99.14**	**99.11**	SCSC
[1]	92.15	91.40	92.35	91.87	SCSC (noise)
ours	**97.07**	**96.81**	**96.85**	**96.83**	SCSC (noise)

However, three different methods were tested on different datasets. So, we implement the method of [1] on SCSC dataset to make our work be more reasonable. Table 3 shows that [1]'s result is still lower than that of ToneNet. We also added the gaussian noise to the data of SCSC when extracting features. And we can see the result from the last line of Table 3 that the performance of ToneNet is still higher than other method and it has achieved 97.07% of accuracy, 96.81% of precision, 96.85% of recall and 96.83% of f1-score.

Table 4: *Confusion matrix of ToneNet's test results without gaussian noise.*

Tone	T1	T2	T3	T4
T1	536	0	1	0
T2	0	394	2	0
T3	2	7	451	1
T4	0	1	2	515

In order to further analyze the performance of ToneNet, we counted the confusion matrix of ToneNet's test result. From Table 4 we can see that ToneNet has a strong ability to distinguish each type of tone.

5. Conclusions

The task of monosyllabic tone classification is an important part of speech evaluation. Based on the convolutional neural network and multi-layer perceptron, this paper proposes a ToneNet model which is suitable for the classification of monosyllabic tones in Mandarin Chinese. We trained and tested ToneNet with the dataset of SCSC and analyzed the focus of ToneNet. ToneNet has achieved 99.16% of accuracy, 99.08% of precision, 99.14% of recall and 99.11% of f1-score. Moreover, ToneNet has achieved 97.07% of accuracy, 96.81% of precision, 96.85% of recall and 96.83% of f1-score in the condition of gaussian noise.

6. Acknowledgements

The research of this paper was supported by the Fundamental Research Funds for the Central Universities (2017XNG1749), and the project of Research on Evaluation Technique of Children's Mandarin Speech Training (HG1711-1).

7. References

[1] Charles Chen, Razvan C. Bunescu, Li Xu, Chang Liu, "Tone Classification in Mandarin Chinese Using Convolutional Neural Networks," Interspeech, 2016.

[2] G.A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in Proc. of HLT, 2006.

[3] X. Lei, M. Siu, M. Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," in Proc. of Interspeech, 2006.

[4] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extractingand composing robust features with denoising autoencoders," in Proceedings of the 25th International Conference on Machine Learning. ACM, 2008, pp. 1096–1103.

[5] F. Chen, L. L. Wong, and Y. Hu, "Effects of lexical tone contour on Mandarin sentence intelligibility," Journal of Speech, Language and Hearing Research, vol. 57, no. 1, pp. 338–345, 2014.

[6] Convolutional Neural Networks (LeNet) - DeepLearning 0.1 documentation. DeepLearning 0.1. LISA Lab. [31 August 2013].

[7] Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25, pp. 1106–1114, 2012.

[8] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proc. International Conference on Learning Representations http://arxiv.org/abs/1409.1556 (2014).

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of CVPR, pp. 770–778, 2016. arxiv.org/abs/1512.03385.

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of ICML, pp 448–456, 2015. jmlr.org/proceedings/papers/v37/ioffe15.pdf.

[11] Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. Proc. Int. Conf. Mach. Learn. 807–814 (2010).

[12] L. Bottou. Stochastic gradient tricks. In Neural Networks, Tricks of the Trade, Reloaded, pp. 430–445. Springer, 2012.

[13] O. Kalinli, "Tone and pitch accent classification using auditory attention cues," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 5208–5211.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In arXiv:1610.02391v3, 2017. 1, 2.

[15] Brian McFee, Colin Raffel, Dawen Liang, Daniel. PW Ellis, MattMcVicar, Eric Battenberg, Oriol Nieto, "librosa: Audio and music signal analysis in python", Proceedings of the 14th Python in Science Conference, 2015.

[16] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4868–4872.