Dialect Species Recognition Based on WaveNet

Xiao Li[1,a], Zhe Liu[1,b] and Zihao Chen[1,c]

[1]*Faculy of Information Technology Beijing University of Technology* Beijing, China
a. 382052461@qq.com, b. 806005126@qq.com, c. chenzihao000@qq.com

Keywords: Dialect species recognition, waveNet, dilated convolution, feature extraction.

Abstract: This paper presents a dialect species recogn-ition technology based on dilated convolutional neural networks. Dialects as a unique national culture, have rich cultural connotations. If Chinese dialects are to be identified systematically, the dialects must first be classified and summarized to determine the species of the dialects. Three dialects are selected to construct the corpus, and digitize and preprocess the audio data. The features of Mel Frequency Cepstrum Coefficients (MFCC) and FBank are extracted. WaveNet-based convolutional neural network structure is trained to save the best model. The integration of the residual network (ResNet) makes the expressiveness of the network proportional to its depth. Organize classification labels and save the mapping between label classifications and dialects. The experimental results show that the improved accuracy of WaveNet can improve the recognition accuracy to more than 90%, which can be used in dialect accent identification and other fields.

1. Introduction

In the existing technology, most of the speech systems only have the ability to recognize standard Mandarin, which has caused inconvenience to many dialect speakers, and has formed obstacles in communication in scenarios such as normal chat and customer service. Language recognition is to determine the language type of a speech automatically, as a front-end processing technology for related language applications. language recognition applies in multilingual speech recognition, information retrieval and services, instant messaging systems, public security fire protection systems, and machine translation field. Speech recognition is similar to speaker recognition in that we are trying to extract information about the entire utterance rather than the content of a particular word[1].

Guided by China's policy of vigorously promoting standard Mandarin, a local language environment with Mandarin as the standard Chinese language has long been formed. In this context, various dialects have formed. The social education role of dialects in regional cultural promotion and cultural characteristics inheritance should be carried forward[2]. We need to use technology to protect some dialects that are fading out of people's attention.

The language recognition system previously represented by GMM-HMM as an acoustic model has gradually evolved into an acoustic model based on deep learning[3]. In 2016, iFLYTEK proposed a fully sequential convolutional neural network (DFCNN) to directly convert a speech sentence into an image as input, Through more convolutional layers and pooling Combination of layers to model the entire speech[4]. CNN is an alternative type of neural network that can be used to reduce spectral variations and model spectral correlations which exist in signals. Since speech signals exhibit both of these properties, CNNs are a more effective model for speech compared to Deep Neural Networks (DNNs)[5]. CNN's unique convolution structure, combined with dilated causal convolution, can also handle long-term sequence information dependence. WaveNet is an auto-regressive deep generation model, which uses dilated causal convolutional neural networks to model and solve the problem of long-term sequence dependence. Use maximum likelihood criteria to optimize model parameters. Based on the convolutional neural network, the residual network and skip connection are fused[6]. Based on the WaveNet technology, this paper improves it and integrates a network layer that is more suitable for dialect species identification, thereby speeding up the calculation speed, improving parallelism and recognition accuracy.

2. Speech Signal Preprocessing

As a non-stationary time-varying signal, the speech signal contains various information in the speech. Before language analysis, the main thing we need to do is to effectively and efficiently extract and represent the characteristic information carried by the speech signal. The waveform of the voice signal changes continuously in time, and its potential amplitude is an analog signal. To process this signal on a sedigital signal system, we need to A/D convert it to a digital signal for storage and computer processing. The process of A/D conversion requires two operations: sampling and quantization. The sampling frequency is selected as 16kHz. According to Nyquist sampling theory, it is selected as twice the signal bandwidth to ensure high-quality signals, and no information is lost during the sampling process. The quantization step is placed after sampling, and the actual 12-bit quantization is commonly used[7].

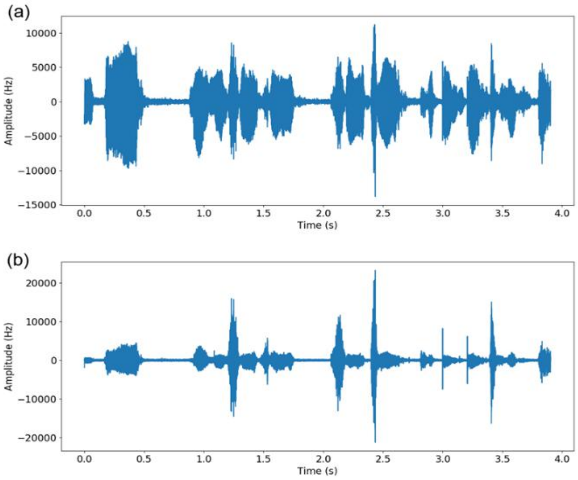

Figure 1: Pre-emphasis.

In order to remove the lip radiation to the voice signal and improve the high-frequency resolution, we need to pre-emphasis the high-frequency part of the digitized voice signal. Taking a Shanghai dialect as an example, “Figure 1” (a) shows the time domain waveform of the speech signal before pre-emphasis. “Figure 1”(b) shows after pre-emphasis. Generally, we use high-pass filters to achieve high-frequency processing of speech signals. First-order FIR high-pass digital filter(α= 0.97) [9]:

$$H(z) = 1 - \alpha z^{-1}, \quad 0.9 < \alpha < 1 \tag{1}$$

The result after pre-emphasis processing is:

$$y(n) = x(n) - \alpha x(n-1), \quad 0.9 < \alpha < 1 \qquad (2)$$

Where x (n) is the speech sample value at n time.

The human vocal organs have inertial activity, and it can be considered that the speech signal is approximately unchanged in a short period of time, corresponding to the short-term stability of the speech signal. The speech signal is divided into short segments, each segment is 10ms~30ms, which is a frame. For the smoothness between frames, the frame will use the frame overlap to add the concept of frame shift, so that the voice characteristics can be observed The continuity of changes and the correlation between adjacent frames ensure more accurate analysis[9]. Frames can be selected using rectangular windows, Hamming windows:

$$\boldsymbol{w[n]} = \begin{cases} \mathbf{0.54 - 0.46} \boldsymbol{cos}\left(\frac{\boldsymbol{2\pi n}}{\boldsymbol{N-1}}\right), \mathbf{0} \le \boldsymbol{n} \le \boldsymbol{N} - \mathbf{1} \\ \mathbf{0}, \qquad\qquad\qquad\qquad \boldsymbol{else} \end{cases} \qquad (3)$$

3. Dialect Species Recognition

3.1.Wavenet

3.1.1. Causal Convolution

WaveNet is a new generation model that can be operated directly on the original audio waveform. It was proposed by Google DeepMind Lab in 2016.The joint probability of a waveform **x** = *{x*1*, . . . , xT }*is factored into the product of multiple conditional probabilities[6]. For the additional given input h, there is a transition from formula (4) to (5).

$$p(x) = \prod_{t=1}^{T} p(x_t | x_{1,\ldots,} x_{t-1}) \qquad (4)$$

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_{1,\ldots,} x_{t-1}, h) \qquad (5)$$

3.1.2. Gated Activation Units

WaveNet use the same gated activation unit as used in the gated PixelCNN. The activation function:

$$z = tanh\,(W_{f,k} * x) \odot \sigma(W_{g,k} * x) \qquad (6)$$

where σ is the sigmoid non-linearity, k is the number of the layer, ⊙is the element-wise product and ∗ is the convolution operator[10].

3.1.3. Dilation Convolution

Dilated convolutions support exponentially expanding receptive fields without losing resolution or coverage. The receptive field grows exponentially while the number of parameters grows linearly, the dilation is doubled for every layer up to a limit and then repeated[11].

e.g. 1, 2, 4, . . . , 16, 1, 2, 4, . . . , 16, 1, 2, 4, . . . , 16.

3.2. Network Structure

Considering that WaveNet is mainly used in the field of speech synthesis, based on this, the potential problems of this model are analyzed in depth to build a model with excellent performance. Using the extracted features as the input layer and activating them with the Relu activation function. The shape of the output feature map does not change. The activation operation makes the neural network learn a non-linear mapping.

It is batch-standardized and integrated into the residual block[12], and then back-propagation (BP) algorithm is used for discriminative training. The pooling layer uses GlobalAveragePooling to reduce the dimensionality of the data. Keep softmax as the last layer due to classification issues. When classifying, it is not necessary to consider the previous classification situation and the adjacent context information, then abandon causal convolution.

"Figure 2" shows the improved model structure. Taking FBank features as an example, 40 filters are selected to form the filter bank, we can extract the acoustic characteristics of each frame of audio signal as the input of the network. The input layer is followed by a one-dimensional convolution layer with 1 × 1 kernel size filter. The BN layer and the activation layer are arranged behind each convolution layer. The main input and output shape changes are shown. The omitted parts are the activation layer and the batch normalization layer, their input and output shapes are the same.

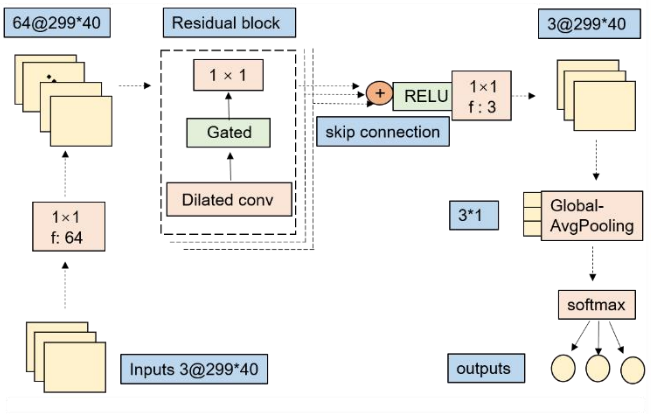

Figure 2: Network structure mode with Fbank.

4. Experiment

4.1. Experiment Procedure

The data set we used is from 2018 iFLYTEK AI Challenge competition, we select three of them as the data basis, and finds that the training is overfitting using crude experiments. Using a dropout operation and discarding some neurons may solve this problem. It is necessary to increase the relevant data set. The Chinese dialects in the Common Voice data set were selected and formatted for conversion to the .wav.

After framing and hanming window, when using the method in the python_speech_feature library to extract Mel filter-bank features, the number of filter banks is set to 40 to obtain 40-dim feature vectors. The extracted MFCC feature vector is 13-dim by default. The MFCC calculation process is shown in “Figure 3”[13]

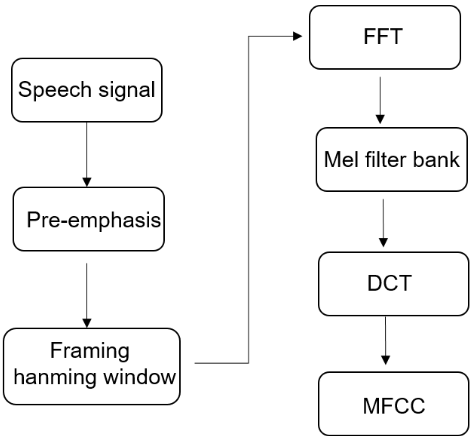

Figure 3: MFCC Feature Extraction.

We can add its first-order difference and second-order difference to get the inverted 39-dimensional stitching feature vector, which shows the dynamic characteristics. Normalize the feature data in the training and test sets according to formula (7): where the mean is μ_i and the variance σ_i.

$$\tilde{o}_i^m = \frac{o_i^m - \mu_i}{\sigma_i} \tag{7}$$

Taking the above features as the input of the WaveNet-based neural network structure, and performing μ-law coding on it, the huge computational load of the neural network is reduced, and the entire network becomes more lightweight, then it`s time to train the model.

4.2. Experiment Result and Analysis

In “Figure 4”, the original WaveNet is used to implement speech recognition. In the first seven epoch of experiments, the extracted FBank features can achieve higher accuracy than the MFCC features. From the eighth round, the accuracy obtained by the two features fluctuates slightly.

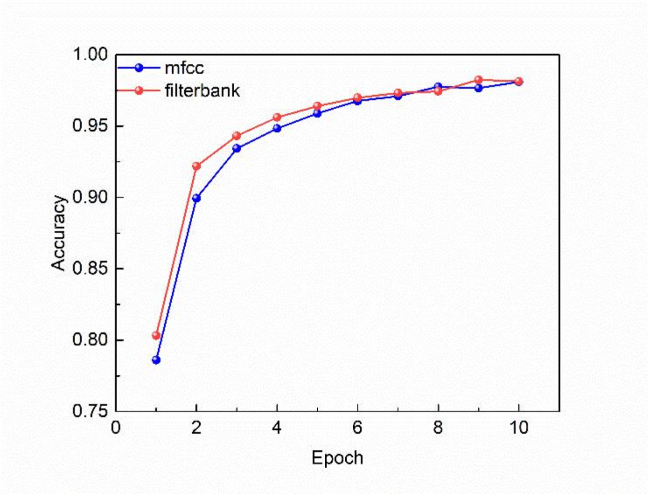

Figure 4: MFCC vs fbank feature accuracy.

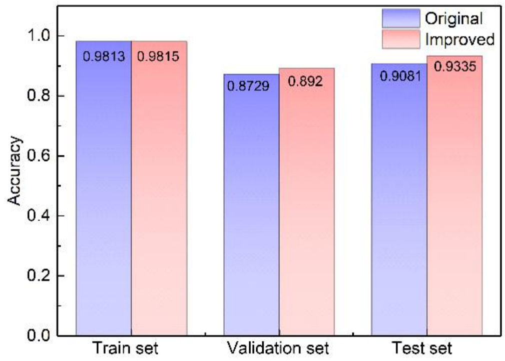

Figure 5: Original model vs improved model with fbank.

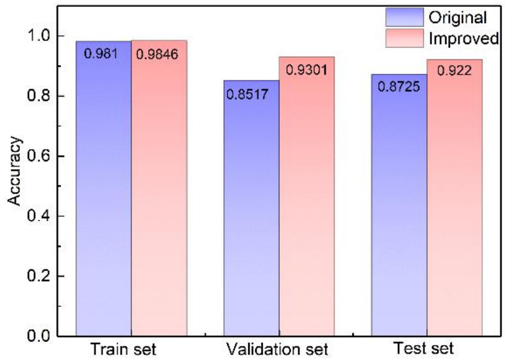

Figure 6: Original model vs improved model with MFCC.

"Figure 5" and "Figure 6" show the accuracy of FBank and MFCC as input for the two models under different data sets.Original Model uses causal dilated convolution in WaveNet. Improved Model only uses dilated convolution. The accuracy of the MFCC test set under the basic model is 87.25%, the Fbank is 90.81%, and the entire training process of MFCC lasts 21159s. The accuracy of the test set of MFCC under the improved model is 92.20%. Fbank is 93.35%, MFCC takes 18592s, which is about half an hour faster; Compared with the WaveNet network, the improved model takes any one of the features as input and can improve the accuracy of the basic model by about 3% ~5%.

5. Conclusions

In order to allow the machine to better recognize our dialects, iFLYTEK and Baidu have also made a lot of efforts to implement dialect protection plans. The network model in this paper can recognize the dialect types of non-specific people. In the future, we hope to use the original audio as a breakthrough from the language types with small dialect differences, and try to extract more representative features of each dialect, and find more Dialect data set in order to train a more general model, thereby improving the recognition efficiency and accuracy, ensuring the scientificity of the experimental results.

References

[1] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, P. A. (2006). Support vector machines for speaker and language recognition. Computer Speech & Language, 20(2-3), 210-229.

[2] Yu Zhang. Research on the Existence Mode of Chinese Dialects in the Mandarin Environment [J]. Course Education Research, 2018 (29): 24.

[3] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. Applied Intelligence, 42(4), 722-737.

[4] Haikun WANG, Jia PAN,Cong LIU. Research development and forecast of automatic speech recognition technologies[J]. Telecommunications Science, 2018, 34(2): 1-11.

[5] Sainath, T. N., Mohamed, A. R., Kingsbury, B., & Ramabhadran, B. (2013, May). Deep convolutional neural networks for LVCSR. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 8614-8618). IEEE.

[6] Aaron van den Oord, Sander Dieleman, Heiga Zeny. WAVENET: A GENERATIVE MODEL FOR RAW AUDIO[J].2016.

[7] Jin Wu. A practical course of speech signal processing [M]. People's Posts and Telecommunications Press: Beijing, 2015.2: 31-45.

[8] Dalmiya C P, Dharun V S, Rajesh K P. An efficient method for Tamil speech recognition using MFCC and DTW for mobile applications[C]//2013 IEEE Conference on Information & Communication Technologies. IEEE, 2013: 1263-1268.

[9] Ittichaichareon C, Suksri S, Yingthawornsuk T. Speech recognition using MFCC[C]//International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012). 2012: 28-29.

[10] Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., & Graves, A. (2016). Conditional image generation with pixelcnn decoders. In Advances in neural information processing systems (pp. 4790-4798).

[11] Yu F , Koltun V . Multi-Scale Context Aggregation by Dilated Convolutions[J]. 2015.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[13] T. Wiatowski and H. Bölcskei, "A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction," in IEEE Transactions on Information Theory, vol. 64, no. 3, pp. 1845-1866, March 2018.