Speech Event Recognition Model for People with Dysarthria Based on Deep Learning

Yi Yang[1], Zipeng Zhang[1, *]

[1]Sun Yueqi College, China University of Mining and Technology, Jiangsu 221008, China

[*] Corresponding Author

Abstract

Dysarthria is a problem faced by many patients with special diseases, which causes speakers to have unclear pronunciation. In order to better understand the speech events expressed by patients with dysarthria, this article proposes a new speech event recognition model based on deep learning. The model takes speech clips as input, uses Gram angle field to retain the original features of the time series, uses Conformer to extract local features and global features of the sequence, and finally uses ResNet as a classification model. Experimental results on the EasyCall corpus data set show that the model proposed in the article has good recognition results.

Keywords

voice event recognition, Gramian Corner Field, Conformer; ResNet.

1. Introduction

In recent years, the number of patients with dysarthria due to certain special diseases[1] such as stroke or cerebral palsy has been increasing. Dysarthria[2] is a language disorder in which patients have slow speech expression and unclear pronunciation due to brain damage. At present, automatic speech recognition(ASR)[3], Voice Conversion (VC)[4]and Voice Event Recognition (Audio Event Recognition, AER)[5] and other research directions are mainly used to solve the problem of communication inconvenience caused by dysarthria. Research in these directions mainly involves the following types of models: nonlinear models (including Gaussian mixture models[6], hidden Markov models[7] and non-negative matrix factorization models[8]), models based on deep neural networks (such as long short-term memory network[9], recurrent neural network[10] and other models). However, these types of models are unable to comprehensively extract various features of the speech sequence, making it difficult to better understand the patient's expression.

For the problems existing in the existing models, this paper uses the Conformer[12] model based on CNN and Transformer[11]model to overcome them. Conformer is a Transformer based on convolution enhancement, which can better obtain global interactive information and local features. In speech recognition systems, speech data are usually represented by concatenated mel-spectral coefficients or perceptual linear prediction coefficients, but this approach does not allow the model to analyze the structural characteristics of time series features. To address this problem, this article uses Gram angle fields[13] to encode time series speech data into images, which can better represent the time dependence and retention of the data. The absolute value of the time relationship.

2. Method

The algorithm introduced in this article uses speech clips as input to model. The model mainly incorporates Gram angle field, Conformer and ResNet[14].

First, the Gram angle field is used to convert the input one-dimensional time series into a Gram matrix, which can not only maintain the time dependence of the sequence, but also retain the absolute value of the time relationship; secondly, Conformer combines the structure of convolution and Transformer Combined, the local features and global features of the sequence are extracted; finally, ResNet is used for classification to obtain the corresponding speech event category. The overall structure of the model is shown in Figure 1.

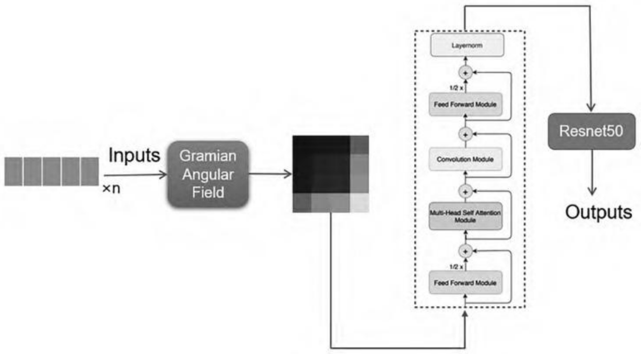

Figure 1: Overall structure of speech event recognition model

2.1. Gram's Corner Field

The processed data are input into the Gram's angle field. The Gram angle field generates a Gram matrix from a one-dimensional time series in a Cartesian coordinate system using trigonometric functions. This method enables data to be classified into time series using deep neural networks. Gram sum angular field(GASF) and Gram difference angular field(GADF) are defined as follows:

$$GASF = [\cos(\phi_i + \phi_j)] = X^{'} \cdot X - \sqrt{I - X^2}^{'} \cdot \sqrt{I - X^2} \tag{1}$$

$$\mathrm{GADF} = [\sin(\phi_i + \phi_j)] = \sqrt{I - X^2}^{'} \cdot X - X^{'} \sqrt{I - X^2} \tag{2}$$

Where:I is the unit row vector [1,2,…,1].

Gram matrices provide a way to preserve time dependencies and maintain absolute time relationships through polar coordinates.

2.2. Conformer

The data processed by Gram's angle field is input into Conformer. Conformer is based on the convolution-enhanced Transformer. It models the local and global dependencies of audio sequences. It integrates Transformer and CNN to build a convolution-enhanced Transformer model and achieves the best results of both.

2.3. Gram's Corner Field

The features extracted by Conformer are input into ResNet, and ResNet is used for classification, and finally the event category is obtained. ResNet proposes a residual learning structure that represents each layer as a residual function of the learning reference layer input to aid the training of deep networks. A large number of experimental results show that residual networks are easier to optimize and gain accuracy from increasing depth.

ResNet solves the degradation problem of accuracy by introducing a deep residual learning structure, and uses residual learning to train each cascade layer.

3. Experiment

3.1. Datasets

The database utilized in this study is the EasyCall corpus, which currently comprises 16,683 speech samples from 21 healthy individuals and 26 individuals with speech impairments. By learning from and analyzing these speech data, the model can identify the speech events expressed in the language of people with dysarthria. Prior to the input of the speech data into the model, preprocessing of the data is required, which includes slicing and normalization.

3.2. Model training settings

Regarding the data part, 70% of the data is used for training and the remaining 30% is used for testing. In the optimizer part, the Adam optimizer[15]is used. The Adam optimizer can not only adapt to sparse gradients, but also alleviate the problem of gradient oscillation. Regarding the adjustment of the learning rate, the CosineAnnealingWarmRestarts function[16]is used, and the cosine annealing plan is used to set the learning rate of each parameter group. Regarding the loss function part, since the model in this article identifies event categories, the cross-entropy loss function is used.

In the computer configuration, the CPU version is E5-2680 v4, the GPU version is NVIDIA RTX A5000, and the PyTorch version is 1.12. The model is trained with the following configuration. The model is trained for 200 epochs. During training, the loss value is calculated and retained for each batch of each epoch. The loss value of each epoch is obtained by averaging the loss values of all batches in an epoch.

3.3. KNN algorithm analyzes in data set

The loss functions during the training process and testing process are shown in Figure 2 and Figure 3. It can be seen from the figure that although the loss value fluctuates during the intermediate training, the loss value gradually decreases as the number of training times increases, and finally tends to steady and stay low.

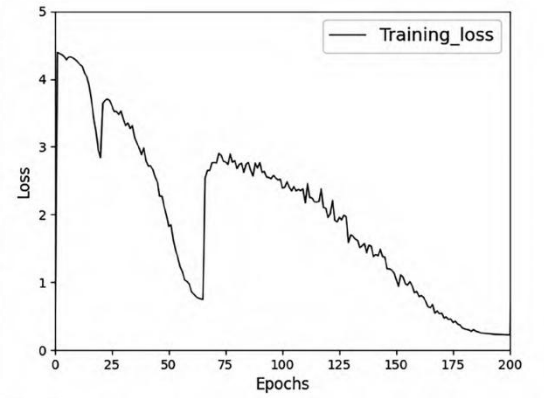

Figure 2: Loss value of training set

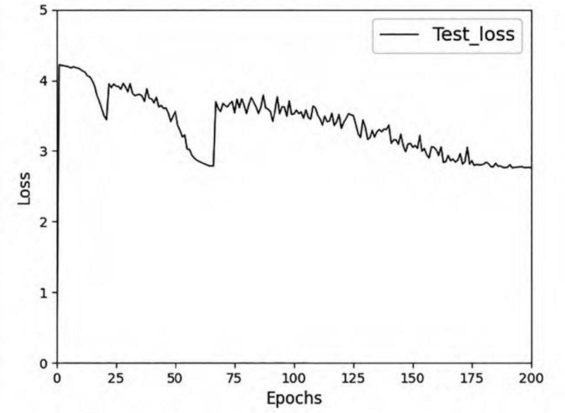

Figure 3: Loss value of test set

During the training and testing processes, the accuracy fluctuated, but the overall trend was upward and gradually stabilized, with the accuracy of the test set reaching 68%.

4. Conclusion

This paper proposes a speech event recognition model based on deep learning: first, the Gram angle field is used to convert the input one-dimensional time series speech data into a Gram matrix; secondly, the local features and global features of the sequence are extracted through the Conformer . Finally, ResNet is used for classification to obtain the corresponding speech event categories, which shows good speech event recognition results on the data set. The innovations of this article mainly include two aspects: using Gram's angle field to process speech data can maintain the time dependence of the data and retain the absolute value of the time relationship of the data; using Conformer can more comprehensively extract speech features and can capture them well. Local features and learning global interactive information.

References

[16]CHEN K C, YEH H W, HANG J Y, et al. A joint-feature learning-based voice conversion system for dysarthric users based on deep learning technology. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019: 1838-1841.

[17]ENDERBY P. Disorders of communication: dysarthria. In: Handbook of clinical neurology, 2013, vol. 110: 273-281.

[18]MAKHIJANI R, SHRAWANKAR U, THAKARE V M. Opportunities & challenges in automatic speech recognition. ArXiv preprint arXiv:1305.2846, 2013.

[19]MOHAMMADI S H, KAIN A. An overview of voice conversion systems. Speech Communication, 2017, 88: 65-82.

[20] KUMAR A, RAJ B. Features and kernels for audio event recognition. ArXiv preprint arXiv:1607.05765, 2016.

[21]VIROLI C, MCLACHLAN G J. Deep Gaussian mixture models. Statistics and Computing, 2019, 29(1): 43-51.

[22] MOR B, GARHWAL S, KUMAR A. A systematic review of hidden Markov models and their applications. Archives of computational methods in engineering, 2021, 28(3): 1429–1448.

[23] LEE D, SEUNG H S. Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, 2000, vol. 13.

[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780.

[25] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization. ArXiv preprint arXiv:1409.2329, 2014.

[26] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. In: Advances in neural information processing systems, 2017, vol. 30.

[27] GULATI A, QIN J, CHIU C C, et al. Conformer: Convolution-augmented transformer for speech recognition. ArXiv preprint arXiv:2005.08100, 2020.

[28] WANG Z, OATES T. Imaging time-series to improve classification and imputation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[29] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.

[30] KINGMA D P, BA J. Adam: A method for stochastic optimization. ArXiv preprint arXiv:1412.6980, 2014.

[31] LOSHCHILOV I, HUTTER F. Sgdr: Stochastic gradient descent with warm restarts. ArXiv preprint arXiv:1608.03983, 2016.