Gsdnet: Gated Self-Supervised Denoising Speech Control Network

Caitong Bai¹, Xiaolong Cui², *, and Ai Li¹
¹Graduate Group, Engineering University of PAP, Xi ’an, China
²Lu Mu Campus, Engineering University of PAP, Xi ’an, China

*Corresponding author: bet_edu@163.com

Abstract. The cost of labeling data remains high, even with the effective implementation of deep neural networks in speech recognition. At the same time, noise still hampers the performance of speech-recognition methods. Thus, it is still challenging to make full use of data sets to enhance the robustness of recognition systems. In this letter, we construct GSDNet, a gated self-supervised denoising speech control network that consists of three parts (a denoising feature-extraction frontend, a speech recognition encoder, and a decoder based on gated convolutional neural networks with self-supervised regression), to provide a low-cost method for training a robust speech recognition system, and we apply it to equipment-control tasks. Finally, the experimental results with the THCH30 and AISHELL data sets for equipment control show that the word error rate is less than 0.2 without a language model.

Keywords: Speech recognition, Self-supervised learning, Gated convolution network.

1. Introduction

Deep neural networks, which have gone through a stage of development from GMM-HMM [1] to DNN-HMM [2] to end-to-end [3] in speech recognition systems, have become mainstream due to their strong generalizability and ability to be embedded in mobile devices with requiring many computational resources. Among these methods, Chan proposed LAS [4] and introduced seq2seq for speech recognition, and Graves proposed CTC [5] to perform real-time CTC translation. Then, introducing an attention mechanism to further strengthen the full use of contextual information, Graves proposed RNN-T before neural transducer, and MoChA [6] is currently being developed. However, the quantity of data required to solve speech-recognition problems is also increasing as the number of system parameters increases. Improving data utilization and reducing training costs have become critical areas of research. Concurrently, noise strongly affects recognition when moving from a laboratory to a real scene, and robust speech recognition based neural networks are rapidly being developed. Currently, there are two mainstream algorithms, namely, designing features [7] and regression algorithms such as GAN, that are used to enhance data and create an adaptive model; for example, Ravanelli proposed a self-supervised multitask network.

Therefore, improving recognition robustness while reducing labeling costs has become a critical problem, and many successful attempts have been made for robust speech recognition, such as deep autoencoders, GANs and auxiliary tasks. However, due to the characteristics of Mandarin speech
signals and the task of controlling equipment requiring a higher robustness, robust Mandarin speech recognition remains a challenging task and must be strengthened. In this letter, GSDNet is developed based on our latest research, using self-supervised training in two stages to solve an auxiliary task that creates a gated neural network consisting of three components: a feature extraction module, an encoder and a phase decoder. The primary innovations of this article include the following: 1) an antinoise robust neural network that uses a deep residual gated neural network and self-supervised training methods to extract acoustic features; and 2) a deep residual gated neural network that achieves a high accuracy with a Mandarin speech-recognition encoder and decoder that are designed for the control of equipment. The rest of this letter is organized as follows. We first formulate the required tasks before describing the composition of GSDNet on III and IV. We then discuss the training analysis in V and the performance evaluation in V compared with some classical algorithms. Finally, in VI, we conclude this letter.

2. Speech Feature Extraction Module

We refer to as the speech feature extraction module of the robust speech recognition network, which consists of four components: the distortion module, which adds noise; the Gated-CNN module, which extracts features and removes noise; the QRNN module; and the adjuster module, which adjusts and enhances the extracted feature.

2.1. Distortion

To simulate noise similar to that in a real environment, several interference methods are applied, where \( p \) represents the probability of occurrence of such a type of interference, which ultimately adds an interference signal to the original signal. The clean speech signal can be formulated as above after noise is added:

\[
\text{Dist}(x) = p \cdot T(x)
\]

\[
T \in \{\text{Tem}, \text{Rever}, \text{Fre}, \text{Add}\}, x \in \mathbb{R}^n
\]

where \( \text{add}() \) adds noise, such as alarms, door knocks, telephone ringing, and television sounds, to the primary signal; \( \text{Fre}() \) is performed by filtering the time signal with band-stop filters; and \( \text{Rever}() \) is introduced by convolving the input signal with a set of 1300 impulse responses derived with the image method.

Figure 1. Model overall architecture diagram
2.2. Gated-CNN
A deep residual gated convolutional network is used to extract the characteristics of the signal after distortion to improve the quality of the extracted features. Then, to increase the network depth while reducing information loss and mitigating gradient disappearance and gradient explosion, the residual structure is introduced to form a residual based mathematical expression of the gated convolutional network, which is:

\[
h_l(X) = (X \ast W + b) \otimes \sigma(X \ast V + c) \\
h_{in}(X) = h_l(X) + h_e(X)
\]

Where \(X\) is a three-dimensional tensor that represents the input tensor in the input layer and the output tensor of the previous layer in the middle layer, and \(W\) represents the convolutional kernel. The construction of the space is described in Fig. 1. Experiments [8] have shown that the gating mechanism can ensure that nonlinear feature extraction and gradient dispersion are reduced, while the calculation speed of the convolutional network is increased.

2.3. QRNN
Because the cyclic neural network can capture the longdistance relationship between the frontend and backend, and to speed up the calculation, QRNN [9] is used as the last layer of the feature extraction module. The mathematical expression of QRNN refers to that in [9], in which each sliding window is used as the input of one QRNN iteration, and each iteration does not rely on the last hidden state so that the network can be operated in parallel.

2.4. Adjuster
In [10], an auxiliary task is used to finalize the improvement of the target tensor and is applied to the adjustment module of Mandarin speech recognition, which is called the adjuster. We build three simple neural networks to fit three artificially extracted features: GFCC, GTN, and LPS. In the first pretraining process, network parameters are updated via the loss function formed by the label.

3. Speech Recognition Encoder and Decoder
The structure of \(f_\theta\) is constructed by the deep residual gated convolution unit. However, the input tensor at this time is the feature extracted by the feature extraction module after pretraining, which is a two-dimensional tensor. Then, half of the channels are used as control gates to restrict the output of the other half of the channels. Each layer of the gated network is connected by a cross-layer residual network to compensate for the loss of information and delay gradient explosion. The overall structure of the decoding network consists of two layers of LSTM network decoders. Their outputs are converted into probability values that propagate through the layers, and the conversion process is as follows:

\[
y_i = \sum_{j=1}^{e^k} \frac{e^{y_j}}{e^j}
\]

Which describes the predicted probability. The tensor output by the LSTM is converted into a probability value output by the softmax layer. Because the output is null, the cluster search algorithm is applied to the algorithm to decode Mandarin characters in order to improve the decoding accuracy. This process calculates the maximum probability of the continuous step under the current pending word as the optimal output. The mathematical expression of the cluster search algorithm [11] is:

\[
\arg \max_y \prod_{i=1}^{l} p(y^{o_i} | x, y^{v_i}, \ldots, y^{o_i-1})
\]
4. Experiments and analysis
We described the network structure and algorithm of GSDNet in the previous sections. Next, we use public and private data sets to train and test the network and compare its results with those of other algorithms.

4.1. Experimental Setup
The training and testing of the model are performed on Facebook’s open-source deep learning framework: PyTorch. To verify the superiority of the proposed algorithm, three open-source speech libraries are used as benchmark test data sets that are easily available: THCH30, AISHELL-1, and STCMDS. To construct an equipment control command data set, we use the exogenous speech synthesis API method and recording method to construct a command data set based on private data. The loss function uses CTC loss, and the optimization function uses Adam. The model recognition result is relative to the real label, which is the real voice content. The recognition result can be converted into a label result after replacement, deletion, and insertion. To complete the recognition efficiency performance test of the constructed control instructions. The decoder consists of a two-layer LSTM network plus a bundle searcher. The number of output words is the output units. The hidden layer dimension is the dictionary dim, which is equal to the dimension of the tag dictionary. The final output tensor is output as a probability through the softmax layer, and then, the beam search algorithm is used to perform a cluster search on the three potential predicted words simultaneously.

4.2. Results and Discussion

![Figure 2. Training result graph](image)

All results of the comparison are shown in Fig. 2 along with the comparison between the algorithm of the extraction module and the superiority of the recognition module. The results show that feature-extraction performance improved after adding various components to the network; these results are confirmed by the experimental comparison between the extracted and traditional features of the Mandarin speech recognition system. The three data sets (THCH30, AISHELL-1, and ST-CMDS) are used as reference data sets to compare the performance of the five algorithms and to demonstrate the accuracy of the residual gated convolutional network in the proposed algorithm. CNN-1024[12] is a speech recognition network architecture based on convolutional neural networks; LSTM-2048[13] is a speech recognition network architecture based on recurrent neural networks; CNN-LSTM[14] is a feature extractor based on the CNN; LSTM is the recognition network architecture of the decoder; SEGGCNN[15] is a recognition network architecture based on the residual gated convolutional neural network; and SEGGCNN-LSTM is the proposed algorithm. We use the word error rate (WER) as the comparison index. Fig. 2 shows the results of the algorithm comparison test and indicates that when
different structures are added, performance optimization is achieved. The gated depth residual architecture has superior performance and is thus applied to equipment control. In engineering applications, a word error rate of 13% has been reached, and voice control can be performed effectively. At the same time, via the addition of the feature extraction network, the accuracy of the Mandarin speech recognition system is improved. Via unsupervised training, the potential value of data samples can be fully explored. The next step explores the problem of knowledge transfer.

5. Conclusions
In this letter, we proposed a self-supervised deep residual convolutional network that achieves good performance on robust speech-recognition tasks, effectively solving the contradiction between increased data costs and increased robustness. We showed that the improved performance of the proposed algorithm stems from the combination of a deep high-capacity model and an augmented self-training method. Finally, we examined the influence of each augmentation model on the preceding accuracy of the model. We observed that the model performance for each sound class is influenced in different ways, suggesting that the model performance could be improved further by applying self-supervised representation features while using the data set sufficiently.

Acknowledgments
This work was supported in part by the National Natural Science Foundation of China under Grant No. U1603261 and by the Natural Science Foundation Program of Xinjiang Province, China, under Program No. 2016D01A080.

References
[1] Juang B H and Rabiner L R 1991 Hidden Markov Models for Speech Recognition Technometrics 33 251–72
[2] Graves A and Schmidhuber J 2005 Framewise phoneme classification with bidirectional LSTM and other neural network architectures Neural Networks 18 602–10
[3] Hinton G, Deng L, Yu D, Dahl G E, Mohamed A, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath T N and Kingsbury B 2012 Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups IEEE Signal Processing Magazine 29 82–97
[4] Chan W, Jaitly N, Le Q and Vinyals O 2016 Listen, attend and spell: A neural network for large vocabulary conversational speech recognition 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp 4960–4
[5] Graves A, Fernández S, Gomez F and Schmidhuber J 2006 Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks Proceedings of the 23rd international conference on Machine learning ICML ’06 (New York, NY, USA: Association for Computing Machinery) pp 369–76
[6] Chiu C-C and Raffel C 2018 MONOTONIC CHUNKWISE ATTENTION 16
[7] Bridle J S and Brown M D 1974 An experimental automatic word recognition system JSRU report 1003 33
[8] Dauphin Y N, Fan A, Auli M and Grangier D 2017 Language modeling with gated convolutional networks international conference on machine learning (PMLR) pp 933–41
[9] Bradbury J, Merity S, Xiong C and Socher R 2016 Quasi-recurrent neural networks arXiv preprint arXiv:1611.01576
[10] Noroozi M and Favaro P 2016 Unsupervised learning of visual representations by solving jigsaw puzzles European Conference on Computer Vision (Springer) pp 69–84
[11] Graves A 2012 Sequence Transduction with Recurrent Neural Networks arXiv:1211.3711 [cs, stat]
[12] Wang Y, Deng X, Pu S and Huang Z 2017 Residual convolutional CTC networks for automatic speech recognition arXiv preprint arXiv:1702.07793
[13] Klakow D and Peters J 2002 Testing the correlation of word error rate and perplexity Speech Communication 38 19–28
[14] Zhao J, Mao X and Chen L 2019 Speech emotion recognition using deep 1D & 2D CNN LSTM networks Biomedical Signal Processing and Control 47 312–23
[15] Pratap V, Hannun A, Xu Q, Cai J, Kahn J, Synnaeve G, Liptchinsky V and Collobert R 2018 wav2letter++: The fastest open-source speech recognition system arXiv preprint arXiv:1812.07625