An End-to-End Model Based on TDNN-BiGRU for Keyword Spotting

Shuzhou Chai, Zhenye Yang, Changsheng Lv, Wei-Qiang Zhang
Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University
Beijing 100084, China
chaisz19@mails.tsinghua.edu.cn, yeezy990511@my.swjtu.edu.cn, lvchangsheng@bupt.edu.cn, wqzhang@tsinghua.edu.cn

***Abstract*—In this paper, we proposed a neural network architecture based on Time-Delay Neural Network (TDNN)-Bidirectional Gated Recurrent Unit (BiGRU) for small-footprint keyword spotting. Our model consists of three parts: TDNN, BiGRU and Attention Mechanism. TDNN models the time information and BiGRU extracts the hidden layer features of the audio. The attention mechanism generates a vector of fixed length with hidden layer features. The system generates the final score through vector linear transformation and softmax function. We explored the step size and unit size of TDNN and two attention mechanisms. Our model has achieved a true positive rate of 99.63% at a 5% false positive rate.**

***Keywords*-TDNN; BiGRU; Keyword Spotting; Attention Mechanism;**

I. Introduction

With the development of artificial intelligence, various intelligent terminal devices have emerged. Voice interaction has become an indispensable part of smart life. Users can use specific words to wake up smart devices, which is widely used today. For example, we can use *"Hey Siri"* to wake up Apple Devices and *"Okay/Hey Google"* on Google Home [1] for voice search. Keyword Spotting (KWS) aims to detect predefined and small-sized keywords in an audio stream.

Classical methods used in KWS include the Keyword/Filler Hidden Markov Model (HMM) [2,3,4] and large vocabulary continuous speech recognition systems (LVCSR) [5,6]. LVCSR requires at least tens of hours training corpus including annotated data and a reliable pronunciation dictionary. However, obtaining such data in practical applications requires a high price. In addition, LVCSR need to generate rich lattices, which require a large amount of computing resources. It is not suitable for mobile terminals with low power and limited performance. An HMM model is trained for each keyword while training one or more filler models HMM for the non-keyword speech segments. At the time of detection, keyword discrimination and localization are realized by Viterbi decoding. According to the HMM topology, a large amount of operations and memory may be caused.

The end-to-end KWS systems have become popular in recent years. Deep KWS [7] introduced a multilayer perceptron as an alternative to the HMM-based approach. Sainath and Parada [8] build on this work and use Convolutional Neural Networks (CNN) to achieve better results. They mentioned that reduced the model footprints is the main motivation for turning to CNNs. Later, feed-forward DNN is replaced by more powerful networks such as recurrent neural networks (RNNs) [9] and residual networks (ResNet) [10]. ResNet does not have a good long-term dependence on voice audio. The RNN directly models on the input features without learning the local structure between successive time series and frequency steps. The Convolutional Recurrent Neural Networks (CRNN) [11] combines RNN and CNN with better performance than RNN or CNN.

GRU is a good variant of the LSTM network. It is simpler than the LSTM network. It is also a very popular network. There are three gate functions in LSTM: input gates, forgetting gates, and output gates to control input values, memory values, and output values. There are only two gates in the GRU model: the update gate and the reset gate. Bidirectional Gated Recurrent Unit (BiGRU) provides bi-directional time series features. Each layer of BiGRU includes a forward pass and a backward pass. The output at each moment is determined by the GRUs of the previous moment in opposite directions. The final output at this moment is generated by weighted the forward pass and backward pass.

In this paper, we propose a method combining TDNN and BiGRU to train parameters and improve the accuracy of the KWS system. We use a TDNN model with the ability to extract context information at the same time. It has multiple layers and each layer has sufficient internal connections to ensure that the network can learn complex nonlinear decision surfaces. Secondly, it can depict the relationship between the frames which do not require precise time alignment of the label. And we use a layer of BiGRU to obtain the sequence characteristics of the data. Besides, we introduce the attention mechanism model to weight the average of features and improve the accuracy of the KWS system. Finally through linear transformation and softmax function, we can output the decision score. Our model

The corresponding author is Wei-Qiang Zhang.

This work was supported by the National Natural Science Foundation of China under Grant No. U1836219 and the National Key R&D Program of China.

Zhenye Yang is with the Southwest Jiaotong University, Chengdu 611756, China, Changsheng Lv is with the Beijing University of Posts and Telecommunications, Beijing 100876, China directly outputs the results of keyword detection without complex searches. In addition, we explored the impact of different networks and corresponding parameters, including average attention and soft attention, the step of TDNN and the unit of TDNN.

II. Model Description

A. End-to-end architecture

We propose to use a combination of TDNN and BiGRU models in small footprint keyword spotting. As depicted in Figure 1, we first extract the Fbank features $\mathbf{x} = (x_1, \cdots, x_T)$ and preprocess the audio of different lengths for each audio. Then we use TDNN to get feature $\mathbf{h} = (h_1, \cdots, h_T)$ that contain more time relationships:

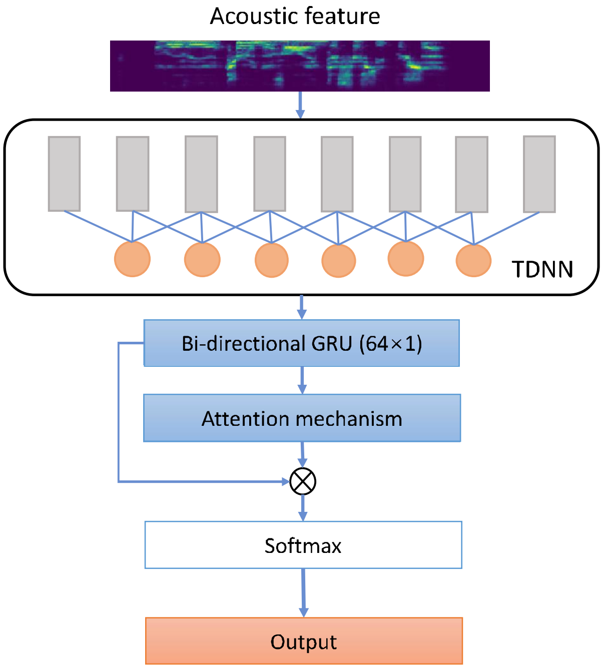

Figure 1. Architecture of our neural networks. TDNN extracts the time information. It generates a vector of fixed length as weighted average of the features of the BiGRU output and the normalized weights of attention mechanism. The Softmax layer outputs the final detection score.

$$\mathbf{h} = TDNN(\mathbf{x}) \tag{1}$$

A set of bidirectional GRUs are utilized to obtain long term dependencies in audios, which results in a feature of high level$\mathbf{g} = (g_1, \cdots, g_t)$:

$$\mathbf{g} = BiGRU(\mathbf{h}) \tag{2}$$

The attention mechanism focuses the model at a position that should be noted, and learns the normalized weight α_t of each frame in the output fetures of the BiGRU:

$$\alpha_t = Attend(\mathbf{g}_t) \tag{3}$$

The output vector d of the model is calculated by weighted average the output **d** features of BiGRU:

$$\mathbf{d} = \sum_{t=1}^{T} \alpha_t \mathbf{g}_t \tag{4}$$

Finally the probability distribution of the result is generated by the linear transformation and the softmax function.We choose cross-entropy as the loss function and rectified linear units (ReLU) [12] as the activation function.

B. Feature Extraction

The Filter-bank (Fbank) feature is selected to extract in the feature extraction model. we generate acoustic features based on 40-dimensional log-filterbank energies computed every 10 ms over a window of 25 ms. Each additional frame of future context adds 10ms of latency to system because of the asymmetry in the input window. To provide the best trade-off between accuracy, latency, and computation [13], we use 5 future frames and 10 past frames. There is no stack of frames based on the GRU system.

C. TDNN Extraction Features

TDNN is a feedforward neural network architecture for modeling time series information. Unlike DNN, the input of the next hidden layer is obtained by splicing the output of the current time and the current time before and after the previous hidden layer. The speech mode[14] suitable for time step can learn a wider time relationship. In this paper, for a segment of speech with length N, each frame has 40-dimensional Fbank characteristics. The input of these time delay units which are 3 steps and 288 units are taken as baselines. These time delay units enter three frame windows spatially, and each unit receives input from the coefficients in three frame windows (through 3*40 weighted connections). Research shows that the 30ms window is sufficient to represent the keyword combination for detection[15].

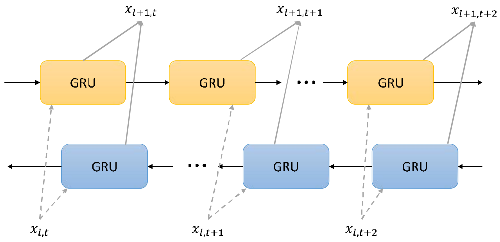

Figure 2. Structure of BiGRU. There are a backward pass a forward pass in every BiGRU layer. Parameter t represents time and l represents the 1 th layer.

D. BiGRU encoder

Long short-term memory (LSTM) is a network of RNNs for modeling long-range dependencies. It is designed to solve the problem of explosion and vanishing gradients. Bidirectional long short-term memory (BiLSTM) combines information of past, current and future time with great success in automatic speech recognition. GRU is a good variant of LSTM. Inspired by the success of BiLSTM, we choose BiGRU as the encoding layer in the model.

There are a forward pass and a backward pass in each BiGRU layer. At each moment, the input provides two

GRUs in opposite directions and the output is determined by the two one-way GRUs. The current hidden layer state of BiGRU is determined by the current input x_t , the output $\overrightarrow{h_{(t-1)}}$ of the forward hidden layer state at $(t-1)$ time, and the output $\overleftarrow{h_{(t-1)}}$ of the reverse hidden layer state:

$$\overrightarrow{h_t} = GRU(x_t, \overrightarrow{h_{(t-1)}}) \tag{5}$$

$$\overleftarrow{h_t} = GRU(x_t, \overleftarrow{h_{(t-1)}}) \tag{6}$$

$$h_t = w_t \overrightarrow{h_t} + v_t \overleftarrow{h_t} + b_t \tag{7}$$

The $GRU()$ function nonlinearly transforms the input audio feature vector and encodes the audio feature vector into the corresponding GRU hidden layer state. W_t represents the forward hidden layer state, $\overrightarrow{h_t}$ corresponding to the BiGRU at time t. V_t represents the weight corresponding to the reverse hidden state $\overleftarrow{h_t}$, t_b represents the offset corresponding to the hidden layer state at time t.

E. Attention mechanism

Attention mechanism makes our model more focused on the part of the speech that contains the keyword and ignores the unrelated part, which is similar to human listening attention.Our attention mechanism provides a better accuracy for KWS. We try both average attention and soft attention. The soft attention is selected to automatically learn how to describe the speech content because of the higher efficiency[16].

Average attention: The attention score α_t is set as the average of T:

$$\alpha_t = \frac{1}{T} \tag{8}$$

Soft attention: Firstly, it learns a scalar score as

$$e_t = v^T \tanh(Wh_t + b) \tag{9}$$

where, h_t is the hidden states. Then soft max is applied to compute the normalized weight as

$$\alpha_t = \frac{\exp(e_t)}{\sum_{j=1}^{T} \exp(e_j)} \tag{10}$$

where, $\alpha_t$stands for the attention score.

III. Experimental Results

A. Data

We train and evaluate the model on an open dataset with the wake-up word “Olivia”(https://drive.google.com/file/d/1m4HIscInvCDbxfU51utMUczcKRZdWv7j/view). The corpus training set consists of 1544 spoken examples, with the keyword “Olivia”. The test set consists of 550 spoken examples, with the keyword “Olivia” and other command statements, such as: Olivia turn the volume to fifty percent.

We also downloaded 15k audio containing noise and speech from the web. We divide the data set into 8k negative test examples and 7k negative train examples.

B. Experiment Setup

Our model is trained with the Adam optimizer [17] and decayed the learning rate after converged from 1×10^{-3} to 1×10^{-4}. The gradient norm clipping to 1 is added to the model and L2 weight decay 1×10^{-5}. We use cross-entropy loss function to reduce the deletion error. The model parameters are trained for 200 epochs with a minibatchsize of 32. The size of the attention mechanism is 100. The number of hidden layer unit in BiGRU is 64.

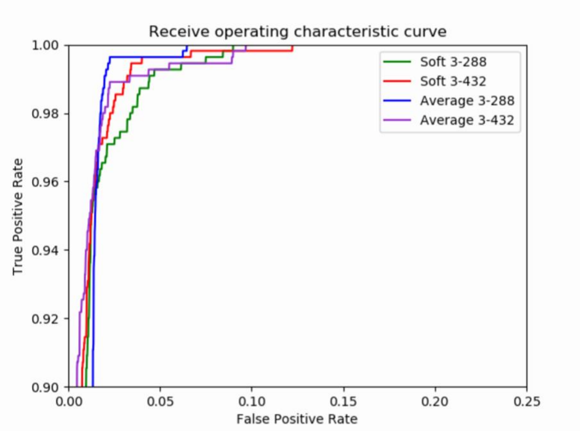

Figure 3. ROCs for soft attention and average attention.

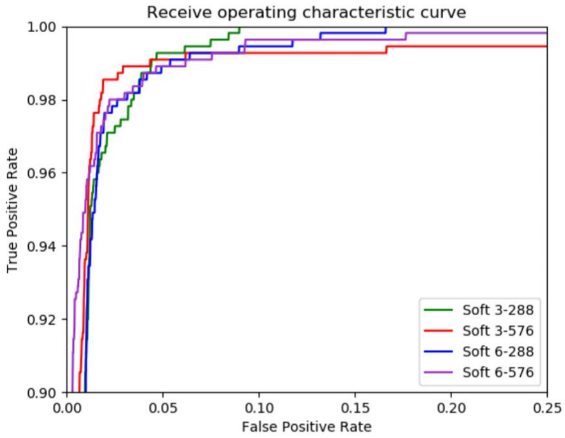

Figure 4. ROCs for different TDNN steps and different TDNN units.

C. Results

We apply the TDNN-BiGRU structure on the ”Olivia” data set. Our KWS baseline system is an end-to-end KWS architecture trained with the Connectionist Temporal Classification loss function in [18]. Our model with 3 steps and 288 units is superior to the baseline. As shown in Table 2, at the false positive rate of 5%, the true positive rate of our model reached 99.63%, which was higher than the baseline of 98.1%. We explored the effect of the number of TDNN units on the accuracy of the model. The results are shown in Figure 4 and Table 1. We observed that as the number of units increases, the true positive rate increases first and then decreases. This indicates that

Table I
PERFORMANCE OF THE PROPOSED MODEL, INCLUDING PRECISION (HIGHER IS BETTER IN ALL INDEXES), RECALL, ACCURACY AND F1. WE COMPARED THE FOLLOWING PARAMETERS: 1) THE AVERAGE ATTENTION MECHANISM AND THE SOFT ATTENTION MECHANISM. 2) TDNN STEP SIZE. 3) TDNN UNITS

Attention	TDNN Step	TDNN Unit	Accuracy	Precision	Recall	F1
Average attention	3	288	97.64%	88.05	76.36	81.79
	3	432	98.41%	96.70	79.82	87.45
	3	576	98.52%	93.56	84.55	88.83
Soft attention	3	288	98.16%	91.39	81.09	85.93
	3	432	98.48%	91.99	85.64	88.70
	3	576	98.59%	92.77	**86.36**	89.45
Soft attention	6	288	98.17%	90.26	82.55	86.23
	6	432	98.47%	92.81	84.55	88.49
	6	576	**98.70%**	**98.48**	82.55	**89.81**

there is an optimal number of units to maximize the true positive rate. However, the recall rate, precision rate and F_1 are almost increasing. Increasing the number of units will convey more hidden information, which is beneficial to the training of network parameters. It can improve the performance of the network.

From Table 1 and Figure 4, we can clearly observe the impact of TDNN step size on model performance. When there are fewer units, the large step size will increase the accuracy, the recall rate and the F1 value of the model. It is because when the step size increases, the TDNN combines more context information to extract features with less information loss. But the amount of calculations in the system will double. When the number of units increases, the recall rate of the model with a smaller step is higher. However, the precision rate of the model with a larger step is higher. When the step is large, the system combines more context information and the information feature extraction is better. Therefore, there is less misjudgment of the examples with keyword when testing and the false positive rate is low.

Figure 3 shows the roc curve of two attention mechanisms. Table 2 compares the true positive rate of different attention mechanisms. As can be observed, when the number of TDNN units is small, the accuracy and true positive rate of the average attention mechanism model are higher than the soft attention mechanism. However, the precision and recall of the average attention mechanism are lower than the soft attention mechanism. When increasing the number of units of TDNN, the true positive rate and the accuracy are similar, and the gap between the precision and the recall is also decreasing. This indicates that the increase of units in TDNN can reduce the impact of the attention mechanism on the results. Therefore, it is possible to consider reducing the amount of calculation by increasing the number of units in the future through using the average attention mechanism.

Table II
COMPARISON OF THE TRUE POSITIVE RATE BETWEEN THE BASELINE SYSTEM TRAINED WITH THE CTC LOSS FUNCTION AND THE TDNN-BIGRU SYSTEM FOR A FIXED FALSE POSITIVE RATE OF 5%.

Model	TDNN Unit	TP
baseline	——	98.10%
Average attention	288	**99.63%**
	432	99.27%
	576	99.13%
Soft attention	288	99.27%
	432	99.63%
	576	99.09%

IV. CONCLUSION

In this paper, we proposed an end-to-end model based on TDNN-BiGRU for keyword spotting. The TDNN-BiGRU-based system has excellent performance on the Olivia dataset compared to the end-to-end model trained with the CTC loss function. Our model consists of three parts: TDNN, BiGRU and attention mechanism. We explored the step size and number of units of TDNN. Experiments show that More units and longer steps can improve the performance of the model. We also explored the impact of the average attention mechanism and the soft attention mechanism on the accuracy of the model. When the number of units is small, the soft attention mechanism works better than the average attention mechanism. When the number of units increases, the two attention mechanisms are not much different. Finally our end-to-end system based on TDNN-BiGRU achieves a true positive rate of 99.63% at a 5% false positive rate.

REFERENCES

[1] J. Schalkwyk, D. Beeferman, F. O. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar and B. Strope "Your word is my command: Google search by voice: A case study," In *Speech Recognition*, Springer, pp. 61–90, 2010.

[2] J. R. Rohlicek, W. Russell, S. Roukos and H. Gish, "Continuous hidden Markov modeling for speaker-independent wordspotting," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 627–630, 1990.

[3] R. C. Rose and D. B. Paul, "AhiddenMarkovmodel based keyword recognition system," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 129–132, 1990,.

[4] J. G. Wilpon, L. G. Miller and P. Modi, "Improvementsandapplications for key word recognition using hidden Markov modeling techniques," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 309–312, 1991.

[5] D. R. Miller, M. Kleber et al, "Rapid and accurate spoken term detection," In *ISCA INTERSPEECH.*, pp. 314–317, 2007.

[6] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p-p. 5244–5247, 2008.

[7] G. Chen, C. Parada and G. Heigold, "Small-footprint keyword spotting using deep neural networks," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4087–4091, 2014.

[8] T. N. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," In *ISCA INTERSPEECH.*, pp. 1478–1482, 2015.

[9] M. Sun, A. Raju, G. Tucker,S. Panchapagesan et al "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting," In *IEEE Spoken Language Technology Workshop (SLT).*, pp. 474–480, 2016.

[10] R. Tang and J. Lin, "Deep residual learning for small-footprint keyword spotting," In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5484–5488, 2018.

[11] S. O. Arik, M. Kliegl, R. Child, J. Hestness, A. Gibiansky, C. Fougner, R. Prenger and A. Coates, "Convolutional recurrent neural networks for small-footprint keyword spotting," In *arXiv preprint arXiv:1703.05390.*, 2017.

[12] V. Nair and G. E. Hinton, "Rectied linear units improve restricted Boltzmann machines," In *Proc. 27th Int. Conf. Mach. Learn. (ICML).*, pp. 1–8, 2010.

[13] X. Lei, A. Senior, A. Gruenstein and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices," In *ISCA INTERSPEECH.*, 2013.

[14] D. E. Rumelhart and J. L. McClelland "Parallel Distributed Processing," In *Parallel distributed processing (Vol. 2).* , Cambridge, MA:: MIT press, 1987.

[15] S. Makino and K. Kido. "Phoneme recognition using time spectrum pattern," In *Speech Commun.*, pp. 225–237, 1986.

[16] F. Chowdhury, Q. Wang, I. L. Moreno and L. Wan. "Attentionbased models for text-dependent speaker verication," In *arXiv preprint arXiv:1710.10470*, 2017.

[17] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization," In *arXiv preprint arXiv:1412.6980*, 2014.

[18] C. Lengerich, A. Hannun. "An end-to-end architecture for keyword spotting and voice activity detection," In *arXiv preprint arXiv:1611.09405*, 2016.