Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 107 (2017) 842 – 847

Procedia
Computer Science

International Congress of Information and Communication Technology (ICICT 2017)

Multilingual Convolutional, Long Short-Term Memory, Deep Neural Networks for Low Resource Speech Recognition

Danish bukhari[a], Yutian Wang[a], Hui Wang[a,*]

Key Laboratory of Media Audio & Video (Communication University of China), Ministry of Education

** Corresponding author：hwang@cuc.edu.cn Tel : +8615801362563*

Abstract

Stand-alone and the combined model of Convolutional Neural networks (CNNs) and Long Short-Term Memory (LSTM) and Deep neural Networks (DNNs) have shown great improvements in a variety of Speech Recognition tasks. In this paper we also combined these networks but in this paper we used them for multilingual speech recognition, for the prediction and correction (PAC) architecture, in order to calculate the state probability. Our proposed model is known as PAC-MCLDNN. In this paper, we present experiment results for multilingual training on AP16-OLR task. Furthermore, cross-lingual model transfer and multitask learning for under resourced languages such as Uyghur and Vietnam are also performed which further improved the recognition results.

Keywords: Multilingual CLDNN, LSTM, CNN, Cross-lingual

1. Introduction

Human perception on understanding what other person says depends upon their own assumption at the next utterance and the conformation of the utterance after it has been uttered. This kind of behavior in human speech recognition has been observed in [10] named as prediction, adaptation and correction (PAC).

Deep neural networks (DNN) [5][6][7][8] have overcome the pervious techniques of HMM/GMM [1][2][3][4] in multilingual speech recognition. Recently, Long short-term memory recurrent neural networks (LSTM-RNNs) [10] and Convolutional neural networks (CNNs) [9] have shown quite a lot of improvements on the multilingual speech recognition task. A combined model for all of these three techniques is shown in [11][12]. None of these jointly trained models used the multilingual data for acoustic model training using the PAC architecture. This work is different in a way that a combined model of CNN-RNN-DNN is used for multilingual speech recognition and also favorable for low resource languages.

Peer-review under responsibility of the scientific committee of the 7th International Congress of Information and Communication Technology
doi:10.1016/j.procs.2017.03.179

Prediction and Correction (PAC) previously used the LSTM RNN and DNN technique to predict the posterior probability by using the stack bottleneck (BN) features from the prediction DNN and used it as an input to the correction DNN [10]. In our work the difference is that the input features are multiple languages provided to the prediction frame to the convolutional neural network. The convolutional layers are stacked with 2 fully connected layers which are further stacked with 2 LSTM layers and later on with multilingual deep neural network layers (MDNN). In order to reduce the dimensions of the last fully connected (FC) layer we used a linear layer followed by [11] and add it as an input to the LSTM layer. The prediction information which is the hidden layer to the correction frame was taken from both the MDNN layers and FC CNN layers. Both the observations were observed and reported in section 4. The correction frame only comprises of the 2 FC CNN along with 2 LSTM layers and MDNN layers with softmax layer at the end of the MDNN layer.

Multitasking technique is applied to acoustic modeling at several occasions [13][14]. The related languages are jointly trained in order to improve the recognition of the target language. The target language should be similar to the related languages for the better accuracy of the system. Previously DNN shared layers are used to improve the accuracy of the target language but no one never used the combined model to improve the accuracy of the system. In this paper, in order to reduce the computational load we adopt the multitasking technique from [10] and shared the hidden layers of the model by keeping the output layers distinct.

As in [9], IARPA-Babel corpus is used entirely focusing on the low resource languages. For our case we used AP16-OLR corpus [15] particularly focusing on multilingual speech recognition tasks. We use Uyghur and Vietnam as our target language to improve the accuracy. The reason of choosing Uyghur as a target language is because it is considered very close to Oriental languages. To our best knowledge for the first time Uyghur language is considered for multi-tasking and multilingual speech recognition.

Uyghur is the southeastern Turkic language which is spoken by ten million people in China and the neighboring countries such as Kazakhstan, Kirghizstan [23]. It is influenced primarily by Persian and Arabic and recently by Mandarin Chinese and Russian

The rest of the paper is structured in a way that in section 2 we gives us the combined multilingual model. Section 3 shows our PAC-MCLDNN architecture. Section 4 shows Experimental setup and Section 5 explains our results followed by conclusion and references.

2. Combined Multilingual Model

The network we use here is a combination of Convolutional, long short term memory and deep neural network in the multilingual framework known as MCLDNN. The MCLDNN model is adopted from the single language input featured CLDNN model in [11].

2.1. Deep shared DNN and CNN

Convolutional neural networks after being widely used in computer vision [17][18] made their way towards speech recognition [19][20]. Our model is adopted from the multilingual VBX network defined earlier in [9]. The difference is that in our work two untied FC layers are used combined with the convolutional layer (CV). Frames of input features along with the contextual vectors are applied as an input to the network. Each frame is 40 dimensional log-mel feature and the kernel size is set to 3*3. The stride is set as similar to the pooling size. We use convolutions to reduce the size of the feature maps hence the padding is applied in the highest layers of the network. The weights and biases for all the languages are not the same. They are all concatenated in the fully connected layers. For the MCLDNN model these both FC layers act as the multilingual shared hidden layers. For just the multilingual convolutional network, we untie the FC layers except the last two layers and combine the last two layers with the convolutional layers with max-pooling after every two convolutional layer.

Another difference is that we concatenated the LSTM layers with the FC layers of CNN. As mentioned in earlier work [10] that two layers of LSTM give better performance. We also stick with the same and used two layers of LSTM. The framework of LSTM is followed from [22].

In the end, the output of the LSTM is passed to the MDNN layers. The MDNN from [21] having 1024 hidden units, shows that multilingual training can give an additional gain which tends to be larger when the amount of data is