Multi-task Learning in Prediction and Correction for Low Resource Speech Recognition

Danish Bukhari(✉), Jiangyan Yi, Zhengqi Wen, Bin Liu, and Jianhua Tao

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{Danishbukhari, jiangyan.yi, zqwen, liubin, jhtao}@nlpr.ia.ac.cn

Abstract. In this paper we investigate the performance of Multitask learning (MTL) for the combined model of Convolutional, Long Short-Term Memory and Deep neural Networks (CLDNN) for low resource speech recognition tasks. We trained the multilingual CNN model followed by the MTL using the DNN layers. In the MTL framework the grapheme models are used along with the phone models in the shared hidden layers of deep neural network in order to calculate the state probability. We experimented with universal phone set (UPS) and universal grapheme set (UGS) in the DNN framework and a combination of both UPS and UGS for further accuracy of the overall system. The combined model is implemented on Prediction and Correction (PAC) model making it a multilingual PAC-MTL-CLDNN architecture. We evaluated the improvements on AP16-OLR task and using our proposed model we get 1.8% improvement on Vietnam and 2.5% improvement on Uyghur over the baseline PAC model and MDNN system. We also evaluated that extra grapheme modeling task is still efficient with one hour of training data to get 2.1% improvement on Uyghur over the baseline MDNN system making it highly beneficial for zero resource languages.

Keywords: MTL · Multilingual speech recognition
Human computer interaction · Uyghur first section

1 Introduction

It is believed that humans can only hear the sound of a spoken language when it is heard along with other graphemes, the lexical contexts and its resemblance or difference from other languages.

Deep neural networks (DNN) [5–8] have overcome the previous techniques of HMM/GMM [1–4] in multilingual speech recognition. Recently, Long short-term memory recurrent neural networks (LSTM-RNNs) [30] and Convolutional neural networks (CNNs) [10] 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 [11–13]. Among them [13] proposed a multitask learning (MTL) approach to construct a static decoding network encoding the multiple context-dependent state

J. Tao et al. (Eds.): NCMMSC 2017, CCIS 807, pp. 80–88, 2018.
https://doi.org/10.1007/978-981-10-8111-8_8

inventories from the distinct acoustic models. Our MTL combined model is adopted from [8] in which the difference is that our model performs the MTL on DNN and we combined it with PAC-CLDNN but [8] just performs the MTL on the shared hidden layers of DNN.

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 [11, 30]. In our work the difference from [30] is that we concatenated the multilingual CNN model with the PAC model to improve the estimation of the phonetic models of a low-resource language by learning other related task(s) together in the DNN layer. If each task shares the same inputs and the respective internal representation, we jointly learn the related tasks so that we can improve the generalization performance of each specific task. In our proposed method it is the mapping between the ordinary and the supplementary tasks in the MTL framework. Furthermore, if a number of low resource languages are to be learned together, we derive a UPS among the languages and use the UPS learning as an additional task in the learning of the multilingual phonetic models. The UPS learning not only implicitly encodes an indirect mapping among the phones of all the involved languages, but also serves as a regularizer for the learning of the phonetic models of each language [8].

Multitask learning is an approach driven from machine learning to improve the overall performance of the learning tasks by jointly learning multiple related tasks together. MTL has been applied successfully in many speech, language, image and vision tasks with the use of neural network (NN) because the hidden layers of an NN naturally capture learning knowledge that can be readily transferred or shared across multiple tasks. For example, [14] applies MTL on a single convolutional neural network to produce state of the art performance for several language processing predictions; [15] improves intent classification in goal oriented human-machine spoken dialog systems which is particularly successful when the amount of labeled training data is limited; in [16], the MTL approach is used to perform multi-label learning in an image annotation application.

As in [30], IARPA-Babel corpus is used entirely focusing on low resource languages. For our case we used AP16-OLR corpus [17] particularly focusing on multilingual speech recognition tasks. We use Uyghur and Vietnam as our target language. The reason of choosing Uyghur as a target language is because it has resemblances with the Oriental languages. To our best knowledge multi-tasking and multilingual speech recognition techniques are applied to Uyghur language.

Uyghur is the southeastern Turkic language which is spoken by ten million people in China and the neighboring countries such as Kazakhstan, Kirghizstan [18]. 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 Sect. 2 shows the combined PAC-MTL-CLDNN architecture. Section 3 shows the experimental setup. Section 3.3 shows the results and the evaluation of the tasks. Section 4 is followed by conclusion and references.

2 PAC-MTL-CLDNN Combined Architecture

2.1 Model Structure

As for the overall architecture of our PAC-MTL-CLDNN model, shown in Fig. 1, we adopted the PAC model from [30]. The major difference from [30] is that in our work inside each prediction and correction frame we use MTL-CLDNN model.

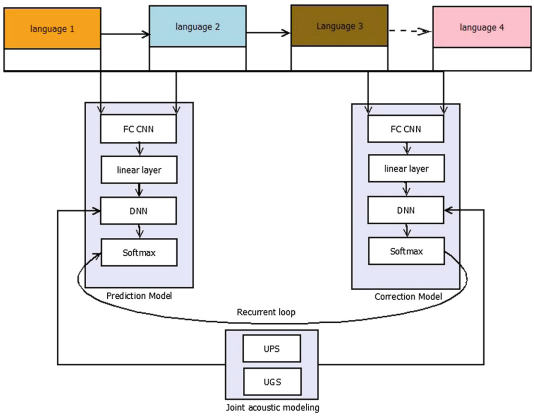

Fig. 1. Overview of PAC-MTL-CLDNN (UPS+UGS) architecture

The correction MTL-CLDNN calculates the state posterior probability [30]. Similar input features are used for prediction MTL-CLDNN. The FC layer of the correction MTL-CLDNN model depends on the FC layer of the prediction MTL-CLDNN model that creates the recurrent loop. The contextual window size is adopted from [30] and they are also set to 10 for the correction MTL-CLDNN and 1 for the prediction MTL-CLDNN. As in [30], the frame cross-entropy (CE) criterion is used. As proposed in [19] for the prediction MTL-CLDNN we used the phoneme label for prediction targets.

2.2 Multilingual CNN Model

Convolutional neural networks after being widely used in computer vision [20, 21] made their way towards speech recognition [22, 23]. Our model is adopted from the multilingual VBX network defined earlier in [10]. The difference is that we use two untied FC layers and combine it 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. With the help of convolutions we reduced the size of the feature maps and 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. In the multilingual CNN framework these both FC layers act

as the multilingual shared hidden layers. 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 from [30] is that we concatenated the LSTM layers with the FC layers of CNN. The framework of LSTM is followed from [24]. As mentioned in earlier work [30] that two layers of LSTM give better performance. We also stick with the same and used two layers of LSTM.

2.3 Multitask Learning DNN Model

The output of the LSTM is passed to the multilingual MTLDNN layers. Our model for DNN is adopted from [8]. The difference from [8] is that the input is a linear layer concatenated with the FC layers of the CNN framework. Second difference is that the FC layers are already the phone models so we experimented with the grapheme model of the specific language to get the evaluation of that language. As in [8] we also used the phone models along with grapheme models as the supplementary task to make a MTL framework. We also created a universal grapheme set (UGS) and a universal phone set (UPS) by taking the unions of the grapheme sets of all the languages which are under investigation. UGS and UPS for the Uyghur language were generated from [28].

2.4 Multi-scale Features

Our aim is to add more information from all multiple languages and use it for further processing without increasing the computation cost. In order to fulfill this need we create different strides on the input window with the help of down sampling. This process is only required at the first conv layer. The parameters are small for the rest of the conv layers so this technique is not required at the other steps.

As for the combination of the conv layers with the DNN layers we add a linear layer to reduce the parameters. The addition of linear layer is seen in [11] but in that it concatenates CNN with LSTM but in our case it is used to combine shared CNN with DNN layers.

2.5 Joint Acoustic Modeling with UPS and UGS

We propose to use a set of universal phone/grapheme (UPS/UGS) as a supplementary learning task along with the phone model training of multiple low resource languages. From the Optimization point of view UPS is used as a regularizer for the phonetic modeling of all the involved languages. From the language point of view it will let the multilingual MTL-DNN to encode a mapping among the phones of all the languages. Trigrapheme models seem to outperform the triphone models with the smaller amount of data. But this performance disappears when full training set was used. This finding helps us to support the use of UGS as an additional task in MTL-DNN framework.

3 Experiment

3.1 Database and Task

AP16-OL7 database comprises of seven different languages from East, Northeast, and Southeast Asia with the main focus on the multilingual Speech Recognition of the Oriental languages [25]. The database is a collaboration between the center of speech and language technologies (CSLT) at the Tsinghua University and Speechocean. For our evaluation we are going to consider 10 h of training set and approximately one hour of speech from the full training set of each language. All are reading style, recordings from the mobile phones with sampling rate as 16 kHz and sample size as 16 bits. We also keep it into consideration that the number of utterances for each speaker remains the same. In addition to that we used THUYG-20 database [26] for Uyghur training. From this database we selected 10 h of training data approx. All are sampled at 16 kHz with sampling size of 16 bits.

3.2 Setup

The proposed training methods were evaluated on two low resource languages i.e. Vietnam and Uyghur. The evaluation shown in Table 1 comprises of complete 10 h of speech data. The configuration of the multilingual CNN is written in Sect. 2. As for the configurations of the DNN model we used 3 hidden layers and 2000 nodes per layer and were trained from 9 consecutive frames. The weights of the hidden layers were initialized by unsupervised pre-training a deep belief network (DBN) of the same architecture [27]. The DBN was configured with the stacking of RBM layers on top of each other and the training was performed layer by layer.

Table 1. Shows the WER% of multilingual systems trained on the 10 h of training data.

System	Vietnam	Uyghur
MDNN	8.9	7.6
MCNN	7.7	6.5
MCDNN	7.5	6.3
PAC-MCLDNN	7.4	5.8
PAC-MTL-MCLDNN-UGS	7.6	5.6
PAC-MTL-MCLDNN-UPS	7.3	5.3
PAC-MTL-MCLDNN (UPS+UGS)	7.1	5.1

During the pre-training stage, mini-batch size was kept steady at 128 (input vectors) with the momentum of 0.5 employed at the beginning which was then increased to 0.9 after 5 iterations. After pre-training, a softmax layer was placed on top of the DBN to get the final model of DNN. This DNN is now a feed forward MLP which is further trained with stochastic gradient descent (SGD). The DNN framework was fine-tuned with a learning rate of 0.02 and was halved with the passage of time due to the performance gain at 0.5%.

The output layer in the MTL-DNN consists of two separate softmax layers one for grapheme and other one from phonemes. For each training sample, two error signals one from each tasks softmax layer were propagated back to the hidden layers. The learning rate remains the same for the output layer but for the input we set it to half.

Two different sets of experiments were performed with the difference in the size of the training data so that we can get the evaluation on the limited amount of data.

3.3 Results and Evaluation

This section presents the experimental results of our study. We trained phone based standalone models of MCNN and MDNN with different initializations. MDCNN models and a combined PAC-MCLDNN model with the concatenation of LSTM/RNN are also trained with the same configurations. After that we trained another set of experiments in which we turn by turn add UPS and UGS as an additional task to the combined model. This addition was in the DNN layers forming a MTL framework. These UPS and UGS were again combined in the proposed network of PAC-MTL-MCLDNN making it PAC-MTL-MCLDNN (UPS+UGS).

We modified the DNN framework by including the universal grapheme set (UGS) as the modeling units. We simply take the unions of all the graphemes involved in these observational languages. As shown in Table 2 the performance drops when we observed with 1 h of training data. We see a relative improvement of 2.4% in the PAC-MTL-MCLDNN (UPS+UGS) framework. Another observation is that UGS performs well with Uyghur language when we have small amount of training data. UGS seems to be useful method for Uyghur language. It shows us that it's a better solution for low-resource language ASR.

Table 2. Shows the WER% of multilingual systems trained on 1 h of small training data.

System	Vietnam	Uyghur
MDNN	10.2	8.9
MCNN	8.9	7.5
MCDNN	8.7	7.4
PAC-MCLDNN	8.4	7.2
PAC-MTL-MCLDNN-UGS	8.1	6.8
PAC-MTL-MCLDNN-UPS	7.9	6.3
PAC-MTL-MCLDNN (UPS+UGS)	7.8	6.2

MDNN was used as a baseline system for our experiments. We performed the experiments on MCDNN. As we are resourced with other language in AP16-OLR corpus we will take an advantage to improve the low-resource languages by exploiting the relationship between phones from multiple languages via a universal phone set in the MTL framework without directly defining the mapping between them.

Numerous techniques on multilingual ASR derive the International Phonetic Alphabet (IPA) or a compact universal phone set (UPS) which is generated by merging the phones in the IPA with the same ASCII format. During multilingual acoustic

modeling, phones available from different languages with the same UPS phonetic symbol will share their training data. Due to this reason we will unite their phone sets by removing all the duplicates from them to build a UPS. This is the supplementary task along with DNN framework which makes it MTL-DNN network.

In the end we combined both the UGS and UPS as the extra learning task in the MTL-DNN framework. In our joint modeling of UPS and UGS the weights are learned in the output layer of the particular language that are determined by learning the weights in the output layer of UPS and UGS. We saw that reducing the training data to 1 h decreased the improvement in the system. The decrease was 1.3% in the MDNN baseline system, 1.0% in PAC-MCLDNN system and 0.7% in the PAC-MTL-MCLDNN (UPS+UGS) system.

This combined network outperforms from the improved PAC-MCLDNN by 0.6% and from the baseline model of MDNN by 2.4%. This gives us the evaluation that MTL is a powerful learning method when the relationship between the languages inside a single corpus is very strong.

This framework further reduces the WER of the overall model. Consistent performance gain is observed for both the larger and smaller training sets in both the tasks. The results demonstrated that MTL performed well in the DNN framework that works well in the combination system. The generalization effect of MTL-DNN training also gives us the observation that the framework performs better on the unseen data. Hence we may conclude that the extra grapheme modeling task is still very effective with an hour of training data. We conceive that this method is highly beneficial for zero resource languages.

4 Conclusion

We believed that the future ASR systems have many compositional components and recurrent feedbacks and they are able to make predictions, corrections and adaptations by themselves. They can judge the number of speakers and then focus on some specific speaker by removing the background noise and other speakers from it. This framework was proposed just to keep this idea in mind.

In this paper, we propose a number of architectures. One is the improvement to the prediction and correction (PAC) model by the addition of multilingual CNN model making it PAC-MCLDNN network. Another is the addition of MTL in the DNN layers of the PAC-MCLDNN network. We carefully sort out the related tasks and utilize positive relationships among them based on our common knowledge. This MTLDNN framework comprises of UPS/UGS prediction as the supplementary task in the PAC-MCLDNN network leading to a PAC-MTL-MCLDNN-(UPS+UGS) system. Our final model outperforms the MDNN model by 2.5% on the full training data and 2.1% on one hour of data. In particular, we found the involvement of UGS in one hour of data as an improvement in the overall system giving a room for the research in low-resource ASR.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (NSFC) (No. 61403386, No. 61273288, No. 61233009), and the Major Program for the National Social Science Fund of China (13&ZD189).

References

1. Burget, L., Schwarz, P., Agarwal, M., et al.: Multilingual acoustic modelling for speech recognition based on subspace Gaussian mixture models. In: Proceedings of ICASSP, pp. 4334–4337 (2010)
2. Mohan, A., Ghalehjegh, S.H., Rose, R.C.: Dealing with acoustic mismatch for training multilingual subspace Gaussian mixture models for speech recognition. In: Proceedings of ICASSP, pp. 4893–4896 (2012)
3. Lu, L., Ghoshal, A., Renals, S.: Regularized subspace Gaussian mixture models for cross-lingual speech recognition. In: Proceedings of ASRU, pp. 365–370 (2011)
4. Lu, L., Ghoshal, A., Renals, S.: Maximum a posteriori adaptation of subspace Gaussian mixture models for crosslingual speech recognition. In: Proceedings of ICASSP, pp. 4877–4880 (2012)
5. Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y.: Cross language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proceedings of ICASSP (2013)
6. Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J.: Multilingual acoustic models using distributed deep neural networks. In: Proceedings of ICASSP (2013)
7. Devin, M., Dean, J.: Multilingual acoustic models using distributed deep neural networks. In: Proceedings of ICASSP (2013)
8. Ghoshal, A., Swietojanski, P., Renals, S.: Multilingual training of deep neural networks. In: Proceedings of ICASSP (2013)
9. Chen, D., Mak, B.K.-W.: Multitask learning of deep neural networks for low-resource speech recognition. ACM Trans. ASLP **23**, 1172–1183 (2015)
10. Sercu, T., Puhrsch, C., Kingsbury, B., Lecun, Y.: Very deep multilingual convolutional neural networks for LVCSR. In: Proceedings of ICASSP (2016)
11. Zhang, Y., Yu, D., Seltzer, M., Droppo, J.: Speech recognition with prediction-adaptation-correction recurrent neural networks. In: Proceedings of ICASSP (2015)
12. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolutional, long short-term memory, fully connected deep neural networks. In: Proceedings of ICASSP (2015)
13. Deng, L., Platt, J.: Ensemble deep learning for speech recognition. In: Proceedings of Interspeech (2014)
14. Bell, P., Renals, S.: Regularization of context dependent deep neural networks with context independent multitask training. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia (2015)
15. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of ICML 2008, pp. 160–167. ACM (2008)
16. Tur, G.: Multitask learning for spoken language understanding. In: Proceedings of ICASSP, pp. 585–588 (2006)
17. Huang, Y., Wang, W., Wang, L., Tan, T.: Multi-task deep neural network for multi-label learning. In: Proceedings of ICIP, pp. 2897–2900 (2013)

18. Wang, D., Li, L., Tang, D., Chen, Q.: AP16-OL7: a multilingual database for oriental languages and a language recognition baseline. In: Proceedings of APSIPA (2016)
19. Smith Finley, J., Zang, X. (eds.): Language. Education and Uyghur Identity in Urban Xinjiang. Routledge, Abingdon (2015). Social Science
20. Palaz, D., Collobert, R., Magimai.-Doss, M.: End-to-end phoneme sequence recognition using convolutional neural networks. In: Proceedings of IJCNN (2013)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of ICLR (2015)
22. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of NIPS, pp. 1097–1105 (2012)
23. Saon, G., Kuo, H.K., Rennie, S., Picheny, M.: The IBM 2015 english conversational telephone speech recognition system. In: Proceedings of Interspeech (2015)
24. Bi, M., Qian, Y., Yu, K.: Very deep convolutional neural networks for LVCSR. In: Proceedings of Interspeech (2015)
25. Sak, H., Senior, A., Beaufays, F.: Long shortterm memory recurrent neural network architectures for large scale acoustic modeling. In: Proceedings of Interspeech (2014)
26. Rozi, A., Wang, D., Zhang, Z.: An open/free database and Benchmark for Uyghur speaker recognition. In: Proceedings of O-COCOSDA (2015)
27. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
28. Saimaiti, M., Feng, Z.: A syllabification algorithm and syllable statistics of written Uyghur. In: CL (2007)
29. Yu, Z., et al.: Prediction-adaptation-correction recurrent neural networks for low-resource language speech recognition. In: Proceedings of ICASSP (2016)