CSLT TECHNICAL REPORT-20160028 [Saturday 7th January, 2017]

Multi-task Recurrent Model for True Multilingual Speech Recognition

Zhiyuan Tang[1,2]
, Lantian Li[1,2]
and Dong Wang[1,3*]

*Correspondence: wang-dong99@mails.tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Research on multilingual speech recognition remains attractive yet challenging. Recent studies focus on learning shared structures under the multi-task paradigm, in particular a feature sharing structure. This approach has been found effective to improve performance on each *individual* language. However, this approach is only useful when the deployed system supports just one language. In a true multilingual scenario where multiple languages are allowed, performance will be significantly reduced due to the competition among languages in the decoding space.

This paper presents a multi-task recurrent model that involves a multilingual speech recognition (ASR) component and a language recognition (LR) component, and the ASR component is informed of the language information by the LR component, leading to a language-aware recognition. We tested the approach on an English-Chinese bilingual recognition task. The results show that the proposed multi-task recurrent model can improve performance of multilingual recognition systems.

Keywords: Multilingual; Multi-task learning; Recurrent neural network; Speech recognition

1 Introduction

Speech recognition (ASR) technologies develop fast in recent years, partly due to the powerful deep learning approach [1, 2]. An interesting and important task within the ASR research is recognizing multiple languages. One reason that makes the multilingual ASR research attractive is that people from different countries are communicating more frequently today. Another reason is that there are limited resources for most languages, and multilingual techniques may help to improve performance for these low-resource languages.

There has been much work on multilingual ASR, especially with the deep neural architecture. The mostly studied architecture is the feature-shared deep neural network (DNN), where the input and low-level hidden layers are shared across languages, while the top-level layers and the output layer are separated for each language [3, 4, 5]. The insight of this design is that the human languages share some commonality in both acoustic and phonetic layers, and so some signal patterns at some levels of abstraction can be shared.

Despite the brilliant success of the feature-sharing approach, it is only useful for model training, not for decoding. This means that although part of the model

structure is shared, in recognition (decoding), the models are used independently for individual languages, with their own language models. Whenever more than one language are supported, the performance on all the languages will be significantly decreased, due to the inter-language competition in the decoding process. This means that the feature-sharing approach cannot deal with true multilingual ASR, or more precisely, multilingual decoding.

A possible solution to the multilingual decoding problem is to inform the decoder which language it is now processing. By this language information, the multilingual decoding essentially falls back to monolingual decoding and the performance is recovered. However, language recognition is subject to recognition mistakes, and it requires sufficient signal to give a reasonable inference, leading to unacceptable delay. Another possibility is to invoke monolingual decoding for each language, and then decide which result is correct, due to either confidence or a language recognizer. This approach obviously requires more computing resource. In Deepspeech2 [6], English and Chinese can be jointly decoded under the end-to-end learning framework. However, this is based on the fact that the training data for the two languages are both abundant, so that language identities can be learned by the deep structure. This certainly can not be migrated to other low-resource languages, and is difficult to accommodate more languages.

In this paper, we introduce a multi-task recurrent model for multilingual decoding. With this model, the ASR model and the LR model are treated as two components of a unified architecture, where the output of one component is propagated back to the other as extra information. More specifically, the ASR component provides speech information for the LR component to deliver more accurate language information, which in turn helps the ASR component to produce better results. Note that this collaboration among ASR and LR takes places in both model training and inference (decoding).

This model is particularly attractive for multilingual decoding. By this model, the LR component provides language information for the ASR component when decoding an utterance. This language information is produced frame by frame, and becomes more and more accurate when the decoding proceeds. With this information, the decoder becomes more and more confident about which language it is processing, and gradually removes decoding paths in hopeless languages. Note that the multi-task recurrent model was proposed in [7], where we found that it can learn speech and speaker recognition models in a collaborative way. The same idea was also proposed by [8], though it focused on ASR only. This paper tests the approach on an English-Chinese bilingual recognition task.

The rest of the paper is organized as follows: Section 2 describes the model architectures, and Section 3 reports the experiments. The conclusions plus the future work are presented in Section 4.

2 Models

Consider the feature-sharing bilingual ASR. Let x represent the primary input feature, t_1 and t_2 represent the targets for each language respectively, and c is the extra input obtained from other component (LR in our experiments). With the information c, the model estimates the probability $P(t_1|x, c)$ and $P(t_2|x, c)$ respectively,

that makes the decoding of two languages absolutely separate. $P(t|x,c)$ is truly required by multilingual decoding, where t means the targets for both two languages. If we regard the extra input c as a language indicator, the model is language-aware. Note that the language-aware model is a conditional model with the context c as the condition. In contrast, the feature-sharing model, which can be formulated as $P(t_1|x)$ or $P(t_2|x)$, is essentially a marginal model $\sum_c P(t_2|x,c)P(c|x)$ or $\sum_c P(t_2|x,c)P(c|x)$, which are more complex and less effective for listing c.

We refer the bilingual ASR as a single task, with respect to the single task of LR. So $P(t|x,c)$ is what we actually compute with the proposed model jointly training ASR and LR, that indicates the two languages use the same Gaussian Mixture Model (GMM) system for generative modeling, though the two languages still use their own phone sets.

We first describe the single-task baseline model and then multi-task recurrent model as in [7].

2.1 Basic single-task model

The most promising architecture for ASR is the recurrent neural network, especially the long short-term memory (LSTM) [9, 10] for its ability of modeling temporal sequences and their long-range dependencies. The modified LSTM structure proposed in [9] is used. The network structure is shown in Fig. 1.

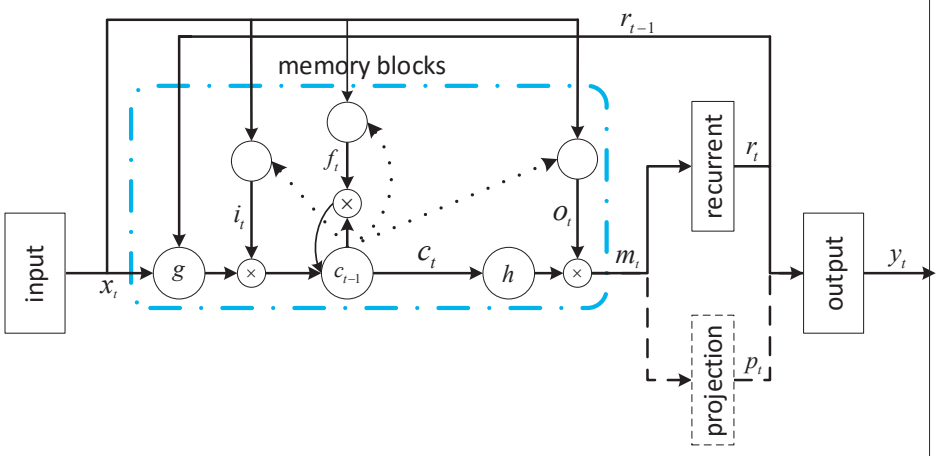

Figure 1 Basic recurrent LSTM model for ASR and LR single-task baselines. The picture is reproduced from [9].

The associated computation is as follows:

$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{ir}r_{t-1} + W_{ic}c_{t-1} + b_i) \\
f_t &= \sigma(W_{fx}x_t + W_{fr}r_{t-1} + W_{fc}c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cr}r_{t-1} + b_c) \\
o_t &= \sigma(W_{ox}x_t + W_{or}r_{t-1} + W_{oc}c_t + b_o) \\
m_t &= o_t \odot h(c_t) \\
r_t &= W_{rm}m_t \\
p_t &= W_{pm}m_t \\
y_t &= W_{yr}r_t + W_{yp}p_t + b_y
\end{aligned}
$$

In the above equations, the W terms denote weight matrices and those associated with cells were set to be diagonal in our implementation. The b terms denote bias vectors. x_t and y_t are the input and output symbols respectively; i_t, f_t, o_t represent respectively the input, forget and output gates; c_t is the cell and m_t is the cell output. r_t and p_t are two output components derived from m_t, where r_t is recurrent and fed to the next time step, while p_t is not recurrent and contributes to the present output only. $\sigma(\cdot)$ is the logistic sigmoid function, and $g(\cdot)$ and $h(\cdot)$ are non-linear activation functions, often chosen to be hyperbolic. $\odot$ denotes the element-wise multiplication.

2.2 Multi-task recurrent model

The basic idea of the multi-task recurrent model is to use the output of one task at the current frame as an auxiliary information to supervise other tasks when processing the next frame. As there are many alternatives that need to be carefully investigated. In this study, we use the recurrent LSTM model following the setting of [7] to build the ASR component and the LR component, as shown in Fig. 2. These two components are identical in structure and accept the same input signal. The only difference is that they are trained with different targets, one for phone discrimination and the other for language discrimination. Most importantly, there are some inter-task recurrent links that combine the two components as a single network, as shown by the dash lines in Fig. 2.

Fig. 2 is one simple example, where the recurrent information is extracted from both the recurrent projection r_t and the nonrecurrent projection p_t, and the information is applied to the non-linear function $g(\cdot)$. We use the superscript a and l to denote the ASR and LR tasks respectively. The computation for ASR can be expressed as follows:

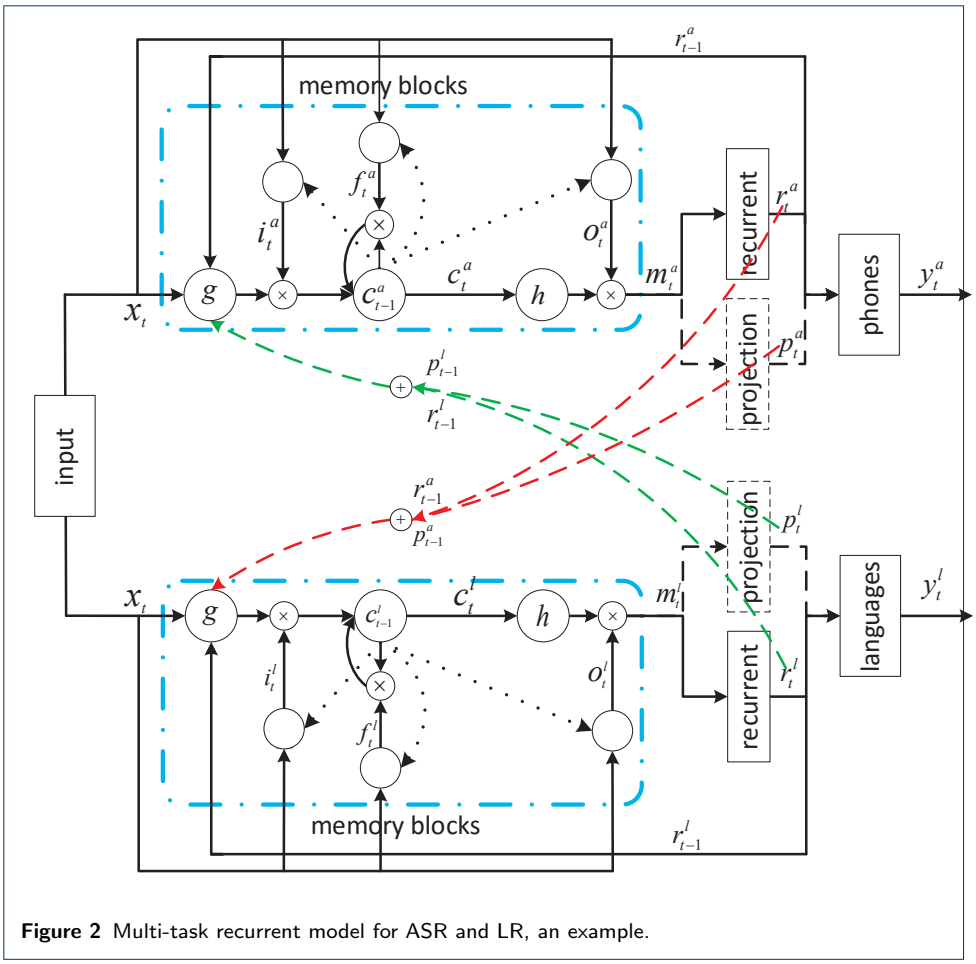

Figure 2 Multi-task recurrent model for ASR and LR, an example.

$$
\begin{aligned}
i_t^a &= \sigma(W_{ix}^a x_t + W_{ir}^a r_{t-1}^a + W_{ic}^a c_{t-1}^a + b_i^a) \\
f_t^a &= \sigma(W_{fx}^a x_t + W_{fr}^a r_{t-1}^a + W_{fc}^a c_{t-1}^a + b_f^a) \\
g_t^a &= g(W_{cx}^a x_t^a + W_{cr}^a r_{t-1}^a + b_c^a + \underline{W_{cr}^{al} r_{t-1}^l + W_{cp}^{al} p_{t-1}^l}) \\
c_t^a &= f_t^a \odot c_{t-1}^a + i_t^a \odot g_t^a \\
o_t^a &= \sigma(W_{ox}^a x_t^a + W_{or}^a r_{t-1}^a + W_{oc}^a c_t^a + b_o^a) \\
m_t^a &= o_t^a \odot h(c_t^a) \\
r_t^a &= W_{rm}^a m_t^a \\
p_t^a &= W_{pm}^a m_t^a \\
y_t^a &= W_{yr}^a r_t^a + W_{yp}^a p_t^a + b_y^a
\end{aligned}
$$

and the computation for LR is as follows:

$$
\begin{aligned}
i_t^l &= \sigma(W_{ix}^l x_t + W_{ir}^l r_{t-1}^l + W_{ic}^l c_{t-1}^l + b_i^l) \\
f_t^l &= \sigma(W_{fx}^l x_t + W_{fr}^l r_{t-1}^l + W_{fc}^l c_{t-1}^l + b_f^l) \\
g_t^l &= g(W_{cx}^l x_t^l + W_{cr}^l r_{t-1}^l + b_c^l + \underline{W_{cr}^{la} r_{t-1}^a + W_{cp}^{la} p_{t-1}^a}) \\
c_t^l &= f_t^l \odot c_{t-1}^l + i_t^l \odot g_t^l \\
o_t^l &= \sigma(W_{ox}^l x_t^l + W_{or}^l r_{t-1}^l + W_{oc}^l c_t^l + b_o^l) \\
m_t^l &= o_t^l \odot h(c_t^l) \\
r_t^l &= W_{rm}^l m_t^l \\
p_t^l &= W_{pm}^l m_t^l \\
y_t^l &= W_{yr}^l r_t^l + W_{yp}^l p_t^l + b_y^l
\end{aligned}
$$

3 Experiments

The proposed method was tested with the Aurora4 and Thchs30 databases labelled with word transcripts. There are 2 language identities, one for English and the other for Chinese. We first present the single-task ASR baseline and then report the multi-task joint training model. All the experiments were conducted with the Kaldi toolkit [11].

3.1 Data

- Training set: This set involves the train sets of Aurora4 and Thchs30. It consists of $17,137$ utterances. This set was used to train the LSTM-based single-task bilingual system and the proposed multi-task recurrent system. The two subsets were also used to train monolingual ASR respectively.
- Test set: This set involves 'eval92' from Aurora4 for English and 'test' from Thchs30 for Chinese. These two sets consist of $4,620$ and $2,495$ utterances and were used to evaluate the performance of ASR for English and Chinese respectively.

3.2 ASR baseline

The ASR system was built largely following the Kaldi WSJ s5 nnet3 recipe, except that we used a single LSTM layer for simplicity. The dimension of the cell was $1,024$, and the dimensions of the recurrent and nonrecurrent projections were set to 256. The target delay was 5 frames. The natural stochastic gradient descent (NSGD) algorithm was employed to train the model [12]. The input feature was the 40-dimensional Fbanks, with a symmetric 2-frame window to splice neighboring frames. The output layer consisted of $6,468$ units, equal to the total number of pdfs in the conventional GMM system that was trained to bootstrap the LSTM model.

The baseline of monolingual ASR is presented in Table 1, where the two languages were trained and decoded separately. Then we present the baseline of bilingually-trained system in Table 2, where a unified GMM system was shared. As for the latter one, we first decoded the two languages with English and Chinese language models (LMs) respectively, denoted as 'mono-LM', and further we merged together the two LMs with a mixture weight of 0.5 using the tool ngram, so both languages can be decoded within a single unified graph built with weighted finite-state transducers, denoted as 'bi-LM'.

Table 1 Monolingual ASR baseline results.

	Englsih	Chinese
WER%	12.40	23.45

Table 2 Bilingual ASR baseline results.

Language Model	Englsih WER%	Chinese WER%
Mono-LM	16.21	23.82
Bi-LM	17.80	23.84

3.3 Multi-task joint training

Due to the flexibility of the multi-task recurrent LSTM structure, it is not possible to evaluate all the configurations. We explored some typical ones in [7] and report the results in Table 3. Note that the last configure, where the recurrent information is fed to all the gates and the non-linear activation $g(\cdot)$, is equal to augmenting the information to the input variable x.

Table 3 Joint training results with Mono-LM.

Feedback Info.		Feedback Input				English WER%	Chinese EER%
r	p	i	f	o	g		
√		√				16.33	23.96
√	√	√				16.27	23.99
√			√			16.15	23.97
√	√		√			16.15	24.01
√				√		16.14	23.90
√	√			√		16.25	23.97
√					√	16.09	23.69
√	√				√	16.34	23.81
√		√	√	√		15.65	23.82
√	√	√	√	√		16.06	23.86
√		√	√	√	√	16.14	23.89
√	√	√	√	√	√	16.32	24.14

Table 4 Joint training results with Bi-LM.

Feedback Info.		Feedback Input				English WER%	Chinese EER%
r	p	i	f	o	g		
√		√				17.81	24.05
√	√	√				17.83	24.03
√			√			17.62	24.02
√	√		√			17.71	23.94
√				√		17.62	23.86
√	√			√		17.69	23.98
√					√	17.54	**23.71**
√	√				√	17.80	23.93
√		√	√	√		**17.21**	23.84
√	√	√	√	√		17.53	23.91
√		√	√	√	√	17.63	23.93
√	√	√	√	√	√	17.93	24.18

From the results shown in Table 3 and 4 decoded with mono-LM and bi-LM respectively, we first observe that the multi-task recurrent model improves the performance of English ASR more than that of Chinese. We attribute this to several reasons. First, the auxiliary component was designed to do language recognition and expected to provide extra language information only, but as the English and Chinese databases are not from the same source, the speech signal involves too

much channel information, that makes the effect of auxiliary language information decrease when channel classification is done at the same time. Moreover, the channel classification was easily achieved by the regular DNN, then the superiority with an additional LR component decays. Second, from the results in Table 2, we find that when using their respective LMs, English gets gains of performance, while that is not obvious for Chinese, even considering monolingual results in Table 1. Results with mono-LM for Chinese in Table 4 were not far away from that of monolingual and bilingual baselines. All imply that a method for improving speech recognition wanting remarkable improvement for this database configuration may not work well. So it's not strange that the performance of Chinese could not be improved much in the enhanced model. Furthermore, we have done another test on part of the train set and all the multi-task recurrent models perform better than the baseline on both English and Chinese, which means the recurrent models overfit the train set extremely, that demonstrates the ability of the proposed model.

We also observe that the multi-task recurrent model still has the potential to exceed the baseline, such as when the recurrent information was extracted from the recurrent projection and fed into the activation function, which led to a better performance for both English and Chinese. We suppose, with many more carefully-designed architectures, the baseline will be surpassed more easily.

4 Conclusions

We report a multi-task recurrent learning architecture for language-aware speech recognition. Primary results of the bilingual ASR experiments on the Aurora4/Thchs30 database demonstrated that the presented method can employ both commonality and diversity of different languages between two languages to some extent by learning ASR and LR models simultaneously. Future work involves using more ideal databases from the same source, developing more suitable architecture for language-aware recurrent training and introducing more than two languages including source-scarce ones.

Acknowledgment

This work was supported by the National Science Foundation of China (NSFC) Project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.
2. D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*, ser. Signals and Communication Technology. Springer, 2015.
3. J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7304–7308.
4. A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7319–7323.

5. G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 8619–8623.
6. D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.
7. Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for speech and speaker recognition," *arXiv preprint arXiv:1603.09643*, 2016.
8. X. Li and X. Wu, "Modeling speaker variability using long short-term memory networks for speech recognition," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2015.
9. H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.
10. H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2014, pp. 338–342.
11. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The kaldi speech recognition toolkit," in *Proceedings of IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.
12. D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.