LETTER *Special Section on Recent Advances in Machine Learning for Spoken Language Processing*

Multi-Task Learning in Deep Neural Networks for Mandarin-English Code-Mixing Speech Recognition

Mengzhe CHEN[†a)], **Jielin PAN**[†], ***Nonmembers***, **Qingwei ZHAO**[†], ***Member***, ***and*** **Yonghong YAN**[†], ***Nonmember***

SUMMARY Multi-task learning in deep neural networks has been proven to be effective for acoustic modeling in speech recognition. In the paper, this technique is applied to Mandarin-English code-mixing recognition. For the primary task of the senone classification, three schemes of the auxiliary tasks are proposed to introduce the language information to networks and improve the prediction of language switching. On the real-world Mandarin-English test corpus in mobile voice search, the proposed schemes enhanced the recognition on both languages and reduced the relative overall error rates by 3.5%, 3.8% and 5.8% respectively.

key words: *multi-task learning, deep neural network, Mandarin-English code mixing, speech recognition*

1. Introduction

Multi-task Learning (MTL) is a machine learning approach where a primary task is learned in parallel with related auxiliary tasks using a shared representation [1]. It has been proposed as a method to improve the generalization of a classifier [2]. In recent works, the MTL using deep neural networks (MTL-DNN) has been applied to the acoustic modeling of speech recognition. This structure improved the training of the network by jointly learning the classifications of phoneme context [3], monophone [4] and articulatory feature [5]. The ability of knowledge integration makes it suitable for many complex tasks, such as low resource recognitions [6], multilingual recognitions [7], recognitions in reverberant environments [8], and so on.

In this paper, a novel use of MTL-DNN in code-mixing speech recognition is proposed. The phenomenon that terms from different languages coexist in an utterance is called code mixing. Code mixing is common in Mandarin Chinese conversations. The statistics from Google Voice Search shows that 10% of the spoken queries obtained from Mainland China contain English words [9]. Even though the two languages (Mandarin Chinese and English) can be recognized well respectively, the challenges are not trivial when they appear together. It is hard for the system to predict the switching of languages in an utterance. On the acoustic model (AM) level, the main reason is the sparsity of code-mixing data. This problem leads to the poor modeling of the modeling units at the switching of languages. In fact, with a data-driven strategy for clustering triphone states to tied triphone states (senones), most of the states of bilingual triphones are merged with the states of monolingual triphones. Here, bilingual triphones mean that the languages of their context phonemes are different from that of their central phonemes, while monolingual triphones mean that the languages of their context and central phonemes are the same. In this way, some information of language switching is lost during the clustering. To solve this problem, three schemes of auxiliary tasks are proposed to introduce the language information into the structure. The first one uses the prediction of phoneme languages which enables the network to learn the discrimination between languages. The second one uses the prediction of phonemes which enables the network to learn the classification of phonemes. The third one combines the first and second tasks, and consists of three parts including the classifications of phonemes and the left and right context phoneme languages. It recovers the language information which is lost during the clustering and enables the network to learn the language context of each phoneme.

The rest of the paper is organized as follows: Section 2 introduces the structure of MTL-DNN and describes auxiliary tasks in detail. The introduction of the recognition system and the experimental results are given in Sect. 3. Finally, Sect. 4 gives the conclusion.

2. Proposed Method

2.1 Multi-Task Learning in Deep Neural Network

The structure of MTL-DNN shares the input layer and the hidden layers. Its output layers for the primary task and the auxiliary task are fully connected to the last hidden layer. If the auxiliary task is chosen properly, this structure can help the primary task be learned better [3]. Another attraction is that when used as an AM in the decoding, the output layer of the auxiliary task will be discarded, so that the model size and usage are the same as the commonly-used single-task DNN.

The structure is given in Fig. 1. x_i represents the ith dimension of the input. D is the total number of dimensions. Given an input vector $\boldsymbol{x}$, the output of the ith nodes of the primary task y_i^p and the auxiliary task y_i^a are derived from softmax function as follows:

Manuscript received January 29, 2016.
Manuscript revised May 11, 2016.
Manuscript publicized July 19, 2016.

[†] The authors are with Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, CAS Beijing, China.
a) E-mail: chenmengzhe@hccl.ioa.ac.cn

DOI: 10.1587/transinf.2016SLL0004

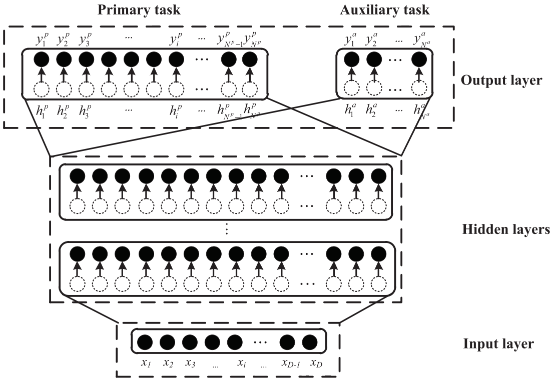
Primary task
Auxiliary task
y_1^p y_2^p y_3^p ... y_i^p ... $y_{N^p-1}^p$ $y_{N^p}^p$
y_1^a y_2^a ... $y_{N^a}^a$
Output layer
h_1^p h_2^p h_3^p ... h_i^p ... $h_{N^p-1}^p$ $h_{N^p}^p$
h_1^a h_2^a ... $h_{N^a}^a$
Hidden layers
Input layer
x_1 x_2 x_3 ... x_i ... x_{D-1} x_D

Fig. 1 The structure of MTL-DNN

$$p(y_i^p|\boldsymbol{x}) = \frac{\exp(h_i^p)}{\sum_{j=1}^{N^p}\exp(h_j^p)}, i \in \{1, 2, \ldots, N^p\} \tag{1}$$

$$p(y_i^a|\boldsymbol{x}) = \frac{\exp(h_i^a)}{\sum_{j=1}^{N^a}\exp(h_j^a)}, i \in \{1, 2, \ldots, N^a\} \tag{2}$$

where h_i^p is the ith activation of the primary task, and N^p is its number of output nodes. Similarly, h_i^a is the ith activation of the auxiliary task, and N^a is its number of output nodes. In this work, the MTL-DNN are trained by minimizing the sum of the cross-entropies of two parts of tasks over all the input features. Thus, the objective function is as follows:

$$\begin{aligned} CE &= \sum_{\boldsymbol{x}}(CE^p + CE^a) \\ &= -\sum_{\boldsymbol{x}}(\sum_{i=1}^{N^p} r_i^p \log p(y_i^p|\boldsymbol{x}) + \alpha \sum_{i=1}^{N^a} r_i^a \log p(y_i^a|\boldsymbol{x})) \end{aligned} \tag{3}$$

where r_i^p and r_i^a are the targets of the ith node for the primary and the auxiliary tasks respectively. α is the weight that controls the proportion of the entropy from the auxiliary task that impacts the back-propagations.

In this work, the primary task is the classification of the senones, and the three schemes of auxiliary tasks will be introduced in Sect. 2.2.

2.2 Schemes of Auxiliary Tasks

The effectiveness of the MTL-DNN depends on the selection of auxiliary tasks. The auxiliary tasks should be related to the primary task and offer extra information to the primary task.

The first scheme uses the phoneme language classification (phnLan task) as the auxiliary task. The added output layer contains three nodes which represent Mandarin Chinese phoneme, English phoneme and non-speech phoneme. To create the targets for each training frame, the targets of senones are mapped down to their corresponding languages according to the central phonemes. This task is designed to give emphasis on the discrimination between the two languages and improve the network ability of the language prediction.

The second scheme uses the task of the prediction of phonemes (phoneme task). It enables the network to learn the discrimination among phonemes and the similarity among the senones which share the same central phonemes. When creating the targets for training, the targets of each frame are mapped down to their central phonemes.

On the basis of the previous schemes, the third one uses the phoneme classification, and the left and right phoneme languages as well (combined task). In this system, three output layers are added to the baseline network. All of them are connected to the last hidden layer of the network like the output layer of the primary task. The CE^a will be separated as follows:

$$CE^a = \beta CE^a_{phoneme} + \frac{\alpha-\beta}{2}CE^a_{phnLan_l} + \frac{\alpha-\beta}{2}CE^a_{phnLan_r} \tag{4}$$

where $CE^a_{phoneme}$, $CE^a_{phnLan_l}$ and $CE^a_{phnLan_r}$ represent the entropies of the phoneme task, the left phoneme language task and the right phoneme language task respectively. The weights of the tasks of phoneme classification and language classification can be assigned to different values for controlling the proportion of the entropies from the two types of tasks. As mentioned above, many bilingual triphones are clustered with monolingual triphones due to the data sparsity of language switching. It will lead to the following problems. Firstly, from the targets, you can not tell the languages of the context. On the other hand, in the data for modeling the senone, the data of the bilingual triphones is overwhelmed by that of the monolingual triphones, so that the senone can hardly present the characteristic of bilingual ones. To avoid losing the information of language switching, the third scheme is designed to bring the language information of each phoneme context to the network, thereby improving the prediction of language switching.

3. Experiments

3.1 Baseline System

Experiments are carried out on Mandarin-English voice search task. The acoustic training data consists of three parts with 236-hour duration. One is 100-hour Mandarin Chinese data: CallHome and CallFriend from LDC database (40 hours), and self-collected in-domain data (60 hours). The second part is 100-hour English data: part of Fisher from LDC (70 hours), and self-collected in-domain data (30 hours). The third part is 36-hour in-domain Mandarin-English data collected by ourselves. The training data of language model (LM) is the text from microblog including 9G mono-Mandarin text and 700M Mandarin-English text. The final LM is obtained by interpolating the microblog part with the in-domain transcripts of AM training.

All experiments use the same network architecture as the baseline system with the exception of the output layer of the auxiliary part. The input to the network is a 572-dimensional vector, which is an 11-frame (5 frames on each

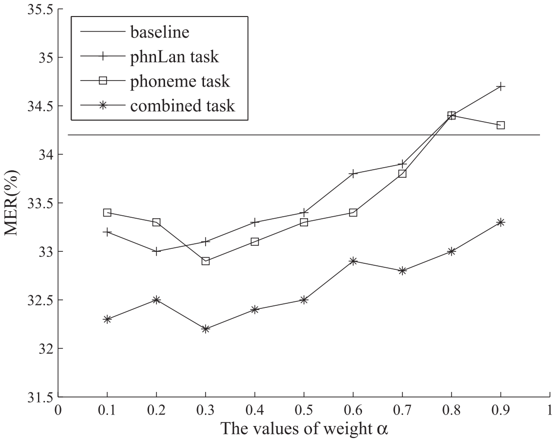

Fig. 2 Performances with different weights

side of the current frame) context window of 52 dimensional features (13-dimension PLP feature along with its first, second and third derivatives). The phoneme set used in the experiments contains 148 phonemes, including 100 Mandarin Chinese phonemes, 39 English phonemes and 9 non-speech phonemes. After clustering, all the triphone states are tied to 5278 senones. Thus, the topology is 572-2048*5-5278. The networks are trained with error back-propagation using cross-entropy objective function. The recognition is carried out by a WFST decoder.

3.2 Recognition Results

The test corpus is collected from the real-world application of mobile voice search. It consists of 1200 Mandarin-English utterances, in which 59%, 19% and 10% sentences have one, two and three English words respectively, and the left 12% sentences have more than three English words.

The performances of the recognition results are measured by mixed error rate (MER) which applies character error rate (CER) for Mandarin Chinese and word error rate (WER) for English. The MER results of the DNNs which are trained with different weights are shown in Fig. 2. Each line corresponds to a scheme of auxiliary task. The performances under different values of α from 0.1 to 0.9 in increments of 0.1 are estimated. In addition, the horizontal line gives the result of the baseline DNN. The results show that the three schemes of MTL-DNNs all achieve better performances than the baseline single-task DNN under appropriate values of α. The scheme of 'phnLan task' obtains the best result when α is 0.2. The best results of 'phoneme task' and 'combined task' both appear when α is 0.3. In the scheme of 'combined task', the parameter β in Eq. (4) can be altered to control the weights of the three parts of auxiliary task. Our experiments indicate that β is not sensitive in terms of changing the system performance significantly, so β is assigned to $\frac{\alpha}{3}$.

To further evaluate the respective recognition results for Mandarin and English, Table 1 reports the CERs for Mandarin Chinese (Man CER) and WERs for English (Eng WER), as well as MERs for the overall performance when α is optimal. Because of the confusion between the two languages, the time mark should be used when aligning recognition results with the reference transcriptions. The insertions, deletions and substitutions are evaluated for each language and summed up for overall evaluation. The results show that with MTL, the performances on both languages are enhanced. The improvement on English is more obvious. The main reason is that if an English word is recognized as Chinese characters, one substitution and several insertions will be added on the errors, thus the WERs of English words are more sensitive to the system performance. The MER results verify the effectiveness of the proposed schemes which obtain 3.5%, 3.8% and 5.8% relative reductions compared with the baseline DNN.

Table 1 Recognition results with the optimal weight

Auxiliary Task	α	Evaluation (%)		
		Man CER	Eng WER	Overall MER
baseline	-	33.1	36.5	34.2
phnLan task	0.2	32.2	34.7	33.0
phoneme task	0.3	32.3	34.2	32.9
combined task	0.3	31.8	33.0	32.2

4. Conclusion

In this paper, the structure of MTL-DNN was adopted in the Mandarin-English LVCSR. In MTL, a better shared internal representation can be learned to improve their generalization performance. Utilizing the advantage of this structure, we introduced the language information into the DNN training. For the primary task of the senone classification, three schemes of the auxiliary tasks were proposed. The experiments on the real-world test corpus showed its effectiveness in enhancing the recognition of code-mixing speech. The recognition improvement on English is more obvious than that on Mandarin Chinese. The best performance was obtained by using 'combined task' of predicting the phoneme and its languages of phoneme context.

This paper aims to offer a method to utilize extra information during training the DNNs for code-mixing speech recognition. The method also can be applied to the code-mixing tasks in other languages.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 61271426, 11504406, 11590770, 11590771, 11590772, 11590773, 11590774), the Strategic Priority Research Program of the Chinese Academy of Sciences (Nos. XDA06030100, XDA06030500, XDA06040603), National 863 Program (No. 2015AA016306), National 973 Program (No. 2013CB329302) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

References

[1] R. Caruana, Multitask Learning, Springer US, 1998.
[2] R.A. Caruana, "Multitask learning: A knowledge-based source of inductive bias," Machine Learning Proceedings, pp.41–48, 1993.
[3] M.L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.6965–6969, 2013.
[4] P. Bell and S. Renals, "Regularization of context-dependent deep neural networks with context-independent multi-task training," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4290–4294, 2015.
[5] H. Zheng, Z. Yang, L. Qiao, J. Li, and W. Liu, "Attribute knowledge integration for speech recognition based on multi-task learning neural networks," 16th Annual Conference of the International Speech Communication Association, pp.543–547, 2015.
[6] D. Chen and K.W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," IEEE ACM Transactions on Audio Speech & Language Processing, vol.23, no.7, pp.1172–1183, 2015.
[7] 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," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7304–7308, 2013.
[8] R. Giri, M.L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5014–5018, 2015.
[9] H.-A. Chang, Y.-H. Sung, B. Strope, and F. Beaufays, "Recognizing english queries in mandarin voice search," 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5016–5019, 2011.