Investigating Multi-task Learning for Automatic Speech Recognition with Code-switching between Mandarin and English

Xiao Song[14], Yuexian Zou[1*], Shilei Huang[2], Shaobin Chen[3], Yi Liu[4]
[1*]ADSPLAB/Intelligent Lab, School of ECE, Peking University, China
[2]Shenzhen Raisound Technologies, Co., Ltd, China
[3]Shenzhen Press Group, China
[4]PKU Shenzhen Institute, China
*E-mail: zouyx@pkusz.edu.cn

Abstract—**This work investigates a Multi-task Learning (MTL-DNN) approach to enhance the performance of Mandarin-English code-switching conversational speech recognition (MECS-CSR). The approach aims at getting a better acoustic model for the primary task by jointly learning two auxiliary tasks together. To overcome the effect of co-articulation at code-switch points, under MTL-DNN, we propose to jointly train two types of Mandarin-English acoustic models according to the choice of acoustic units that describe the salient acoustic and phonetic information for Mandarin. To further make use of language information, we jointly train another acoustic model for language identification (LID) with the two acoustic models under the MTL-DNN. To evaluate the effectiveness of our developed MECS-CSR system, extensive experiments are carried out on a public dataset LDC2015S04. It is noted that our approach does not require other language resources. Compared with the first basic MECS-CSR system [1], Mixed Error Rate (MER) of our proposed approach is relatively reduced by 12.49%. The performance improvement benefits from multi-task learning where the common internal representation is obtained from the auxiliary tasks learning.**

Keywords- Multi-Task Learning; Deep Neural Network; Acoustic Units; Mandarin-English Code-Switching; Speech Recognition.

I. INTRODUCTION

Code-switching is a common linguistic phenomenon that results from different languages coexisting in an utterance among multilingual speakers. Due to the extensive influence of the English language and Mandarin worldwide, the population of bilingual speakers tends to increase and Mandarin-English code-switching in daily conversations becomes more common [2].

Similar to other automatic speech recognition (ASR) systems, the challenges exist in acoustic modeling when we build an ASR system on a code-switch corpus. Especially, due to the sparsity of code-switching data [3], the prediction of the switching languages in an utterance makes it difficult to model the acoustic units at the switching languages. For tackling the problem, [4] [5] investigated speaker adaptation and phone sharing between languages. [6] [7] used monolingual acoustic models in combination with language identification (LID) to recognize code-switching sentences. [1] integrated a LID system into the decoding process by using the multi stream approach. However, even if adding the speaker adaptation information or combining LID, the code-switching problem is still taken as a single task to solve and still tackled faultily.

Recently, multi-task learning with deep neural network (MTL-DNN) has been shown that learning correlated tasks simultaneously can boost the performance of individual tasks. This ability of the knowledge integration makes it suitable for many complex tasks, such as low resource recognition [8], multilingual recognition [9], recognitions in reverberant [10], and so on. In the code-switching speech recognition, a novel use of MTL-DNN was proposed in [3]. It used three schemes of the auxiliary tasks to introduce the language information to MTL-DNN. The three schemes were phoneme language classification, phoneme prediction and combination of the formers. This way provided more language information to enhance the recognition of Mandarin-English code-switching speech. However, there are still not much studies use the MTL-DNN for further enhancing the recognition of code-switching problem on benchmark dataset.

Inspired by the success of single task learning and the correlated tasks learning simultaneously, we propose training initial/final-based and phoneme-based acoustic models together under the MTL-DNN. Obviously, the initial/final modeling and phoneme modeling are highly related tasks. Their joint training does not require additional resources of other languages. The method performs the MECS-CSR by fusing with the hidden layers of MTL-DNN which overcome the challenge of co-articulation effects between phones at code-switching [1] and make full use of the limited training code-switching data.

Furthermore, we study choosing LID as another task, which leverage more language switching information obtained from the MTL-DNN for MECS-CSR task. While the effectiveness of the proposed approach with the further study will be proven for the primary task.

This paper is organized as follows. Section II presents the proposed approach. The introduction of the experimental settings and results are given in Section III. Finally, Section IV gives the conclusion.

II. PROPOSED T-CSR-LID-MTL APPROACH

We propose a T-CSR-LID-MTL approach for simultaneous two types of MECS-CSR tasks and a LID task via MTL-DNN. The framework of the proposed approach is shown in Fig. 1, where the primary task and auxiliary tasks share the hidden representations. How to select single tasks and learn multi-tasks will be shown in details:

978-1-5386-1981-0/17/$31.00 ©2017 IEEE

A. Single Tasks Selecting for the Multi-Task Learning

As shown in the bottom of Fig. 1, there are three tasks: *MECS-CSR-IF-PH*, *MECS-CSR-PH-PH* and *LID*.

Modeling units can be used to describe the salient acoustic and phonetic information for Mandarin Chinese in speech recognition system. Thus, we adopt International Phonetic Alphabet (IPA) to build the universal phone set for the MECS-CSR task. And then, propose two types of single tasks according to the choice of acoustic units for Mandarin Chinese in acoustic model:

1) MECS-CSR-IF-PH task: We choose initial/final for Mandarin and phoneme for English as the acoustic units in the first Mandarin-English acoustic model (*IF-PH*). As known to all, the initial/final as acoustic units can reflect the knowledge and characteristics of Chinese phonetics. And it is widely used to the large vocabulary continues speech recognition system similary to the phoneme for English. We pick the *MECS-CSR-IF-PH* as the primary task for the MTL-DNN.

2) MECS-CSR-PH-PH task: We choose phoneme for both languages as acoustic units in the second Mandarin-English acoustic model (*PH-PH*). Although the phoneme as acoustic units do not have the advantage of initial/final for Mandarin Chinese, the number of acoustic units is less than the initial/final. Thus, in the case of the same amount of training data, the parameters of the phoneme modeling units can be more fully and accurately estimated. We pick the *MECS-CSR-PH-PH* as the first auxiliary task for the MTL-DNN.

The third task, LID, is used to provide more language switching information for the primary task:

3) LID task: We adopt “CH” for Mandarin and “EN” for English as acoustic units in the acoustic model for a simple LID task. [1] intergrated LID information into decoding and obtained 0.3% absolute improvement. In order to make full use of the language information, we utilize LID as the second auxiliary task for the MTL-DNN.

Based on the above discussions, three single tasks are chosen for MTL-DNN.

B. Multi-task Learning Phase for the T-CSR-LID-MTL

As shown in the top of Fig. 1, the *MECS-CSR-IF-PH* (*TASK-S1*), *MECS-CSR-PH-PH* (*TASK-S2*) and *LID* (*TASK-L*) share the hidden representations under the MTL-DNN, which enables these related tasks to be learned together and transfer the knowledge from one task to another task. As a result, the common internal representation thus learned helps the models generalize better for future unseen data.

In the proposed T-CSR-LID-MTL approach, the hidden layers of the MTL-DNN for primary task fuse with ones of the DNN for auxiliary tasks by sharing weights/bias parameters. Here, we assume there are K tasks $T \equiv \{T_1, T_2, \ldots, T_k, \ldots, T_K\}$ to learn. The model parameters are represented by $\Lambda \equiv \{\lambda_0\} \cup \{\lambda_1, \lambda_2, \ldots, \lambda_k, \ldots, \lambda_K\}$, where λ_0 is the model parameters that are shared by all tasks and λ_k is the model parameters associated with task T_k. The parameters Λ are learned ultimately in the MTL-DNN. Without loss of generality, T_1 is taken as the primary task, and the rest are auxiliary tasks. The objective function $CE(D,\Lambda)$ is formulated as the weighted sum of the cross-entropies of K tasks. The optimization of minimizing $CE(D,\Lambda)$ in MTL-DNN can be formulated as follows:

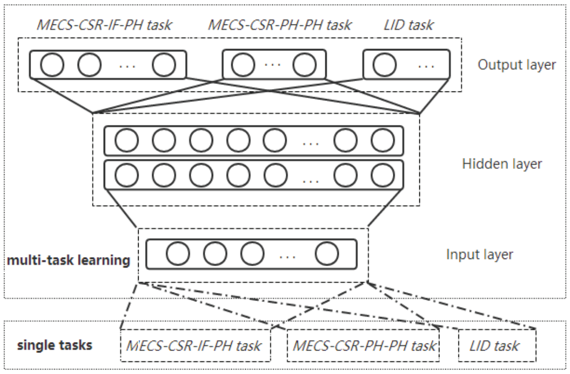

Figure 1. The framework of the proposed T-CSR-LID-MTL approach.

$$CE(D,\Lambda) = \sum_{x \in D} \left(\sum_{k=1}^{K} \beta_k CE_k(x; \lambda_0, \lambda_k) \right) \tag{1}$$

$$CE_k(x) = \sum_{i=1}^{N_k} r_i^k \log p(y_i^k \mid x) \tag{2}$$

$$p(y_i^k \mid x) = \frac{\exp(h_i^k)}{\sum_{j=1}^{N_k} \exp(h_j^k)} \tag{3}$$

where $CE_k(x)$ is the entropy of task T_k. β_k is the task weight of T_k and $\sum_{\mathrm{k=1}}^{\mathrm{K}} \beta_k$. D is the training set and x is one input vector. r_i^k is the target of the i-th node for the task T_k. y_i^k is the output of the i-th node for the task T_k. $p(y_i^k \mid x)$ is the softmax function of task T_k. h_i^k is the i-th activation of the task T_k, and N^k is its number of output nodes.

In order to control the contribution of each task during the process of information sharing, the training objective function CE is designed as follows:

$$CE = \sum_{x \in D} (\beta_{TASK-S1} CE_{TASK-S1}(x) + \beta_{TASK-S2} CE_{TASK-S2}(x) + \beta_{TASK-L} CE_{TASK-L}(x)) \tag{4}$$

where $CE_{TASK\text{-}S1}(x)$, $CE_{TASK\text{-}S2}(x)$ and $CE_{TASK\text{-}L}(x)$ represent the entropy of *TASK-S1*, *TASK-S2* and *TASK-L*, respectively. $\beta_{TASK\text{-}S1}$, $\beta_{TASK\text{-}S2}$ and $\beta_{TASK\text{-}L}$ are the task weight of *TASK-S1*, *TASK-S2* and *TASK-L* ($\beta_{TASK\text{-}S1} \geq \beta_{TASK\text{-}S2} \geq \beta_{TASK\text{-}L}$), which control the tasks of different entropy proportion that impacts the back-propagations to learn different high-level features.

Generally, after training, only the model parameters (λ_0 and λ_1) will be kept and the other parameters ($\lambda_2, \ldots, \lambda_K$) will be discarded [11]. Therefore, the kept λ_0 and λ_1 are learned by the T-CSR-LID-MTL approach for the primary task.

C. The Proposed T-CSR-LID-MTL Approach

The T-CSR-LID-MTL approach is realized as follows:

First, preparing the input features for all tasks: Gaussian Mixture Model (GMM) of the *TASK-S1*, *TASK-S2* and *TASK-L* task are trained respectively. Then, the context dependent decision tree, the audio alignment and the feature transform are adopted from the GMM systems, which are the input of the MTL-DNN.

Second, transferring the knowledge from auxiliary tasks to primary task: the MTL-DNN is trained with the back-propagation through time (BPTT) algorithm. Via the BPTT algorithm, the parameters (weights and bias of each

TABLE I. STATISTICS OF THE DATASETS

	Train Set	**Dev Set**	**Eval Set**	**Total**
Speakers	139	8	10	157
Duration(hours)	53.65	4.00	4.27	61.92
Utterances	48040	1943	2162	52145

TABLE II. CORRESPONDENCE BETWEEN THE TASKS AND THEIR MODELS

Task	**Model**
Single *TASK-S1* (baseline)	STL-DNN-*S1*
Single *TASK-S2*	STL-DNN-*S2*
Single *TASK-L*	STL-DNN-*L*
TASK-S1 fused with *TASK-S2*	MTL-DNN-*S1S2*
TASK-S1 fused with *TASK-S2*, *TASK-L*	MTL-DNN-*S1S2A*

hidden layer) between each task are shared and λ_0 and λ_1 will be learned for the primary task (as description in Section *II.B*).

Third, decoding for the primary task: the final layer is separated for the primary task [11]. Accordingly, the primary task can use its specific language model for decoding individually, which can be addressed efficiently for speech using weighted finite state transducers (WFSTs).

III. EXPERIMENTAL SETTINGS AND RESULTS

A. Dataset

LDC2015S04, a public South East Asia Mandarin-English (SEAME) corpus [12], is employed as our experimental dataset. It contains 63 hours spontaneous Mandarin–English code-switching transcribed speech, which are recorded with the form of unscripted interviews and conversation from 157 Singaporean and Malaysian speakers. About 82% of the transcribed utterances contain one or more intra-sentential code-switches. We divide the corpus into three sets (training, development and evaluation set) approximately the same proportion as [1], where the details are shown in Table I.

B. Setting of the Baseline ASR System

We use the Kaldi speech recognition toolkit [13] to build our baseline for the primary task, where the ASR system employs the single-task learning with Kaldi recipe ‘nnet2’ (STL-DNN). As shown in Table II, the tasks and models are denoted as their abbreviations simply. To build the STL-DNNs, we train the networks using a forced alignment given by GMM systems, which are trained by using Kaldi recipe ‘s5’ (model ‘tri5a’). Accordingly, GMM for each basic task is trained firstly with the standard 13-dimensional Mel-frequency cepstral coefficients (MFCC) features, which are calculated by the 25 ms Hanning windows with 10 ms shift for neighboring frames. Each DNN has four hidden layers with 1024 units per layer. All of DNNs are trained using stochastic gradient descent (SGD) with one weight update per utterance, and the truncated BPTT learning algorithm. At the same time, the learning rate declines from 5×10^{-3} to 5×10^{-4} that is exponentially decayed during training.

In our baseline ASR system for basic single-tasks, N-gram language models are trained with the SRILM toolkit [14] by using the training data transcription. As a result, 3-gram back-off models smoothened using Kneser-Ney discounting are obtained.

C. Settings of the T-CSR-LID-MTL Approach

We implement the proposed approach by using Kaldi

TABLE III. RECOGNITION RESULTS WITH DIFFERENT WEIGHTS OF JOINTING TRAINING THE PRIMIARY TASK AND THE FIRST AUXILIARY TASK UNDER THE MTL-DNN (MTL-DNN-*S1S2*)

Weights of all Tasks ($\beta_{TASK\text{-}S1}$:$\beta_{TASK\text{-}S2}$)	**Evaluation (%)**			
	Pure En. Sen. (CER)	***Pure Ma. Sen. (WER)***	***Pure CS Sen. (MER)***	***Overall (MER)***
1 : 0 (STL-DNN-*S1*)	50.09	39.90	34.21	36.79
0.9 : 0.1	48.90	40.15	32.75	34.95
0.8 : 0.2	48.79	40.01	32.50	**33.71**
0.7 : 0.3	48.55	**36.01**	32.58	34.76
0.6 : 0.4	48.73	39.92	**32.27**	34.59
0.5 : 0.5	**47.63**	39.32	32.55	34.58

TABLE IV. RECOGNITION RESULTS WITH DIFFERENT WEIGHTS OF JOINTING TRAINING THE PRIMIARY TASK AND ALL OF THE AUXILIARY TASKS UNDER THE MTL-DNN (MTL-DNN- *S1S2L*)

Weights of all Tasks ($\beta_{TASK\text{-}S1}$:$\beta_{TASK\text{-}S2}$:$\beta_{TASK\text{-}L}$)	**Evaluation (%)**			
	Pure En. Sen. (CER)	***Pure Ma. Sen. (WER)***	***Pure CS Sen. (MER)***	***Overall (MER)***
1 : 1 : 0	47.63	39.32	32.55	34.58
1 : 0.9 : 0.1	44.36	**36.77**	30.21	**32.03**
1 : 0.8 : 0.2	45.37	37.53	**30.02**	32.16
1 : 0.7 : 0.3	44.63	37.23	30.27	32.21
1 : 0.6 : 0.4	**44.33**	37.42	30.46	32.39
1 : 0.5 : 0.5	45.26	37.20	30.58	32.48

recipe ‘nnet2/train_multilang2.sh’[1].

To fully investigate the performance of the proposed approach, we design the following experiments:

Experiment (I): We joint training *TASK-S1* and *TASK-S2* under the MTL-DNN, which is donated as MTL-DNN-*S1S2* model, shown in Table II.

Experiment (II): We joint training the *TASK-S1*, *TASK-S2* and *TASK-L* under the MTL-DNN, which is donated as MTL-DNN-*S1S2A* model, shown in Table II.

Besides, various task weights have been explored to evaluate the effects of the auxiliary tasks on the final performance.

Every MTL-DNN has the same hidden layers with the same units as the STL-DNNs. As for the training of MTL-DNNs, the model is initialized by the well-trained STL-DNN-*S1*. It is noted that the choice of the initial model doesn’t influence the final results, due to the fact that the final model will converge by two or more iterations. During each iteration, the parameters of each hidden layer between each task are weighted (as description in Section *II.B*) via the BPTT algorithm with the same learning rate of STL-DNNs training.

D. Experimental Results

The performance of the recognition results is measured by mix error rate (MER) which applies character error rate (CER) for Mandarin Chinese and word error rate (WER) for English.

The results of the Experiment (I) with different task weights are shown in Table III. We can observe that the best overall performance relatively reduces MER by 8.37% compared with the baseline when the ratio is 0.8:0.2. When observing single statistics for pure English/Mandarin/code-switching sentences, the best perfor-

[1] https://github.com/xiaosdawn/Kaldi-multi-task/blob/master/egs/wsj/s5/local/online/run_multitask2.sh

TABLE V. RECOGNITION RESULTS WITH THE OPTIMAL WEIGHT

Model	**Evaluation (%)**			
	Pure En. Sen. (CER)	***Pure Ma. Sen. (WER)***	***Pure CS Sen. (MER)***	***Overall (MER)***
IPA based (H.Li[1]-Dev)	-	-	-	37.10
LID+IPA based (H.Li[1]-Dev)	-	-	-	36.60
STL-DNN-*S1* (our baseline)	50.09	39.90	34.21	36.79
MTL-DNN-*S1L* (0.8:0.2)	48.79	40.01	32.75	33.71
MTL-DNN-*S1S2L* (1:0.9:0.1)	**44.36**	**36.77**	**30.21**	**32.03**

mance relatively reduces CER/WER/MER by 4.91%/9.74%/5.67%. This shows that the jointly training initial/final-based and phoneme-based acoustic model benefits learning of the primary task. The improvement on Mandarin Chinese is obvious. The main reason is that our proposed approach aims at the salient acoustic and phonetic information for Mandarin Chinese.

The results of the Experiment (II) with different task weights are shown in Table IV. It is observed that coordinating *TASK-S1*, *TASK-S2* and *TASK-L* simultaneously achieves the best overall performance, which relatively reduces MER by 12.93% compared with the baseline when the ratio is 1:0.9:0.1. When observing single statistics for pure English/Mandarin/code-switching sentences, the best performance relatively reduces CER/WER/MER by 11.50%/7.84%/12.24%. The improvement on pure code-switching sentences is obvious. It can be seen that the LID task provide more language switching information for the primary task under the MTL-DNN.

Finally, results with optimal weight on the proposed T-CSR-LID-MTL approach and the best results on the first MECS-CSR system [1] are shown in Table V. When jointly training *TASK-S1*, *TASK-S2* and *TASK-L*, the best overall performance achieves a relative improvement of 12.93% in terms of MER compared with our baseline. Besides comparing with the first MECS-CSR system on the same corpus [1], the final best result relatively reduces the MER by 12.49%.

IV. CONCLUSION

In this paper, we proposed T-CSR-LID-MTL, a novel approach based on Multi-task Learning with DNN (MTL-DNN) for simultaneous two types of Mandarin-English code-switching conversational speech recognition (MECS-CSR) tasks and a language identification (LID) task. According to the choice of the acoustic units, we choose initial/final for Mandarin and phoneme for English as the acoustic units in the first MECS-CSR, while we choose phoneme for both languages as the acoustic units in the second MECS-CSR. The former task is picked as the primary task, while the latter and the LID task are picked as the auxiliary tasks under the MTL-DNN. The auxiliary tasks transfer knowledge to the primary task by sharing the hidden representations. The experiments were carried out on LDC2015S04 without other resources, and showed the effectiveness of the T-CSR-LID-MTL in enhancing the recognition of Mandarin-English code-switching speech. The best performance was obtained by jointing the primary task with all of the auxiliary tasks simultaneously. A relative improvement of 12.49% were obtained in terms of MER on the evaluation set than the first speech recognition system for Mandarin-English code-switch speech [1] on the development set. In the future, we will focus on the works of language model to build a better MECS-CSR system.

ACKNOWLEDGMENT

This work is partially supported by Key Technologies Research & Development Program of Shenzhen (No: JSGG20150512160434776) and Shenzhen Science & Research projects. (No: JCYJ20160331104524983).

REFERENCES

[1] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, et al., "A first speech recognition system for Mandarin-English code-switch conversational speech," in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012, pp. 4889-4892.

[2] Li, Y., Fung, P.. Code switching language model with translation constraint for mixed language speech recognition. In: Proc. COLING. 2012, p. 1671–1680.

[3] Chen, Mengzhe, et al. "Multi-Task Learning in Deep Neural Networks for Mandarin-English Code-Mixing Speech Recognition." IEICE TRANSACTIONS on Information and Systems 99.10 (2016): 2554-2557.

[4] C. F. Yeh, C. Y. Huang, L. C. Sun, and L. S. Lee, "An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling," in Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on, 2010, pp. 214-219.

[5] S. Yu, S. Zhang, and B. Xu, "Chinese-English bilingual phone modeling for cross-language speech recognition," in Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on, 2004, pp. I-917.

[6] K. Bhuvanagiri and S. Kopparapu, "An approach to mixed language automatic speech recognition," Oriental COCOSDA, Kathmandu, Nepal, 2010.

[7] D.-C. Lyu, R.-Y. Lyu, Y.-c. Chiang, and C.-N. Hsu, "Speech recognition on code-switching among the Chinese dialects," in Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 2006, pp. I-I.

[8] D. Chen and B. K.-W. Mak, "Multitask learning of deep neural networks for low-resource speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, pp. 1172-1183, 2015.

[9] 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 Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013, pp. 7304-7308.

[10] 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," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015, pp. 5014-5018.

[11] D. Chen, B. Mak, C.-C. Leung, and S. Sivadas, "Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 5592-5596.

[12] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–English code-switching speech corpus in South-East Asia: SEAME," Language Resources and Evaluation, vol. 49, pp. 581-600, 2015.

[13] D. Povey, A. Ghoshal, G. Boulianne, O. Glembek, N. Goel, M. Hannemann, et al., The Kaldi Speech Recognition Toolkit, 2012.

[14] A. Stolcke, "SRILM-an extensible language modeling toolkit," in Interspeech, 2002, p. 2002.