Map and Relabel: Towards Almost-Zero Resource Speech Recognition

Ying Shi†, Zhiyuan Tang†, Lantian Li†, Zheling Zhang†, Dong Wang†‡*
† Center for Speech and Language Technologies, Research Institute of Information Technology
Department of Computer Science and Technology, Tsinghua University, China
‡ Beijing National Research Center for Information Science and Technology
Corresponding Author E-mail: wangdong99@mails.tsinghua.edu.cn

***Abstract*—Modern automatic speech recognition (ASR) systems require large amounts of data to train the acoustic model, especially with the state-of-the-art deep neural network (DNN) architecture. Unfortunately, most of the languages in the world have very limited accumulating for data resources, limiting the application of ASR technologies in these languages.**

The state-of-the-art approach to tackle this problem is transfer learning, by which DNNs trained with data of a rich-resource language can be reused by low-resource language systems, in the form of either feature extractor or initial model. This approach, however, still requires several hours of speech, which is still not affordable for many languages. In this study, we present a novel Map and Relabel (MaR) approach that can train ASR systems for new languages with only a few hundred labelled utterances. This approach combines transfer learning and semi-supervised learning in a boosting manner: it firstly trains a simple monophone DNN based on the limited training data, employing the popular transfer learning approach (Map phase); this model is then used to produce pseudo phone labels for a large amount of untranscribed speech (Relabel phase). These pseudo-labelled data are then used to train a full-fledged tri-phone system.

Experiments on Uyghur, a major minority language in the western China, demonstrates that this MaR approach is rather successful: it can train a pretty good ASR Uyghur system by only 500 utterances. This encouraging results indicate that it is possible to quickly construct a reasonable ASR system for any language, and the only effort we need to pay is just labelling several hundred utterances.

I. Introduction

Due to the powerful modeling capability, deep neural networks (DNNs) have become the mainstream model in automatic speech recognition (ASR) [1]. A key ingredient is the availability of large amounts of training data that can be used to learn the complex discriminative functions implemented by DNNs. However, among all the languages in the world, for which the total number is estimated between $5,000$ to $7,000$, only very few can be said rich-resource, e.g., English and Chinese[2][1]. Most of the languages are spoken by a small population and only very limited resources are available, particularly transcribed speech data. This situation hinders the application of ASR technologies in a significant way.

A multitude of research have been conducted to boost performance of DNN-based ASR systems for low-resource languages. A key idea is that human languages share some commonality in both acoustic and phonetic aspects, and so patterns at some levels of abstraction learned by the DNNs can be shared. Inspired by this insight, a multilingual DNN can be trained where the hidden layers of the DNN structure are shared across languages and each language holds its own output layer [3], [4], [5]. This approach further invokes the idea of DNN-based transfer learning, i.e., borrowing data from rich-resource language to enhance modeling for low-resource languages. For example, in the tandem architecture, speech data from rich-resource languages are used to train a bottleneck (BN) feature extractor, which can be used directly to produce features for low-resource languages [6], [7], [8], [9]. In the hybrid architecture, the low-level layers of a DNN trained for a rich-resource language can be excerpted and reused to train a new DNN model for low-resource languages [10], [11]. All these methods can be categorized into the transfer learning paradigm, where the knowledge of the rich-resource language is materialized in the DNN model and then transferred to the new model for the low-resource language [12].

Despite the notable success, this transfer learning approach still requires tens or hundreds of hours of training speech, which is still unaffordable for most minority languages. In this study, we propose a Map and Relabel (MaR) approach to build ASR systems with very limited labelled data, e.g., several hundred utterances. We shall assume the following almost-zero resource condition: we can collect a large amount of speech data, but can only label a few hundred utterances. The limited amount of labelled data is not sufficient to train a reasonable ASR system, even if with transfer learning. This almost-zero condition is typical for many minority languages, especially those that are in the risk of extinction. Providing speech technologies for these languages as quick as possible becomes an urgent task from perspectives of both cultural protection and social benefit. Recently, Chinese government supported a multilingual minority-lingual ASR (M2ASR) project [13], with the goal of developing speech recognition system for five minority languages (Uyghur, Kazak, Tibetan, Mongolia and Kirgiz). This research is part of the M2ASR project, aiming to develop ASR systems for minority languages with very limited speech data, e.g., only 500 labelled utterances.

The MaR approach we developed involves two phases: Map and Relabel. In the Map phase, it borrows a fully-trained DNN for a rich-resource language and uses the limited training data to learn a mapping from the features generated by the

[1]https://en.wikipedia.org/wiki/Language

rich-resource DNN to the phones of the target low-resource language. This is essentially transfer learning. In the Relabel phase, the learned phone mapping are augmented to the rich-resource DNN to form a (target) phone discrimination DNN. This DNN is used to build a simple monophone ASR system, by which the untranscribed speech are pseudo labelled. Finally, the pseudo-labelled speech will be used to train a regular tri-phone system. This relabel phase is essentially a semi-supervised learning.

We verify the MaR approach with a database of Uyghur, a major minority language in the western China. The results demonstrated that this approach is rather effective: it can train an Uyghur ASR system by only 500 labelled utterances, obtaining a performance comparable to the model trained with 50 hours of labelled speech.

The rest of this paper is organized as follow: Section II describes the MaR approach, and Section III describes the experiments. The paper is concluded by Section IV.

II. Map and Relabel (MaR)

The almost-zero resource condition implies that it is not possible to train a full-fledged context-dependent ASR system. We therefore consider a bootstrapping procedure: firstly we build a basic monophone DNN using the limited labelled data, and then use this monophone DNN to label the untranscribed speech. The pseudo-labelled data are then used to build a full-fledged triphone system. The entire procedure can be summarized as 'Map and Relabel', where the Map phase constructs the monophone DNN, and the Relabel phase produces the pseudo labels. Fig. 1 shows this MaR procedure.

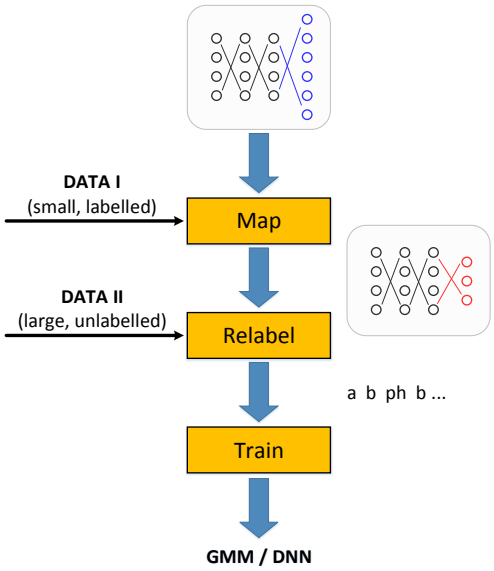

Fig. 1. Diagram of the Map and Relabel procedure.

A. Map

The first phase of MaR is to construct a DNN model that can be used to produce pseudo labels for untranscribed speech. Due to the very limited labelled data, two implementation details are important: (1) the popular transfer learning method that borrows a DNN structure trained for a rich-resource language; (2) a very simple model that involves limited free parameters, so that the limited training data are sufficient. In our work, a large-scale DNN trained with 10k hours of Chinese speech data are borrowed. We reuse the whole structure of the Chinese DNN and except that the targets of the output layer are replaced by the *monophones* of the target language. This equals to borrowing the feature extractor of the Chinese DNN and augments an affine mapping that maps the output of the feature extractor to the phones of the target language. This mapping structure is shown in Fig. 2.

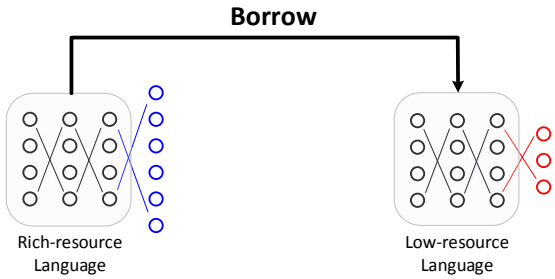

Fig. 2. The Map phase of the MaR approach. The low-level layers of a DNN trained using a rich-resource language (here Chinese) are reused and are augmented with a mapping layer that maps the output of the hidden layers to the phones of the target almost-zero resource language (red circles).

It should be emphasized that the targets of the new mapping-augmented DNN model are phones, rather than state IDs or pdf IDs as usual. This is because the very limited training data prevent us from building complex models. Since the number of target phones is quite small, the mapping layer can be easily trained with the limited data. The resultant DNN is essentially a phone-discrimination model for the target language.

B. Relabel

The second phase of MaR is to label the untranscribed speech by the phone-discrimination DNN. However, directly utilize this DNN to label the speech frame by frame is not reliable: the output is determined by speech signals in a short window, hence rather vulnerable to noise and corruption. A simple approach is to smooth the output using some low-pass filters, but a more powerful approach is to construct a simple ASR decoder that employs both statistical constraints implemented by the HMM architecture and linguistic knowledge implemented by the language model. Since the output of the phone-discrimination DNN is monophone, the decoder can be only a one-state monophone HMM-DNN hybrid system, as shown in Fig. 3. As shown in the next section, this naive decoder can generate pretty good phone labels, in spite of its simplicity.

C. Full training

With the monophone HMM-DNN decoder, we can produce pseudo labels for large volume of speech data. Using the

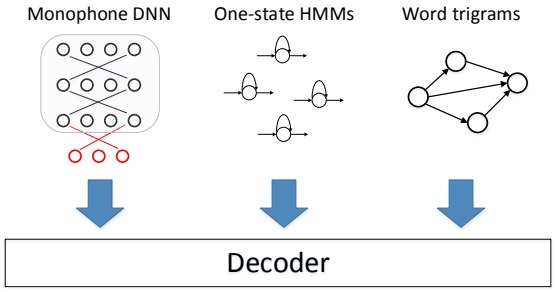

Fig. 3. A simple one-state monophone HMM-DNN decoder.

pseudo labels (phone sequences) as ground truth, a full-fledged triphone system can be trained. According to the semi-supervised learning theory [14], performance can be greatly improved if the pseudo labels are sufficiently accurate. Note that the errors in the pseudo labels may impact differently on different types of systems. We therefore train two systems, one is an HMM-GMM system and the other is a hybrid HMM-DNN system, where the DNN is trained with the criterion of cross entropy (CE).

III. Experiment

A. Data and Model Setting

We use a Uyghur speech database to evaluate the MaR approach offered by the M2ASR project. The entire database involves 50 hours of speech signals from 374 speakers. All the speech signals are collected in the silent office condition, using the same type of carbon microphone. The speaking style is reading. Most of the speakers are college students, and the accent is mostly Urumchi. The database is split into a training set and a test set. The training set involves 50 hours of speech and 348 speakers, and the test set involves 3 hours of speech and 26 speakers. There is no overlap between the two sets in terms of both signals and speakers. More information about the database can be found in the web page of the M2ASR project[2].

To simulate the almost-zero resource condition, 500 utterances of 29 speakers are selected from the training set as the 'labelled set' (LB set, 1.5h in total), which we assume phone-level alignment is available. The rest of the training data are used as 'unlabelled set' (ULB set), which contain 17940 utterances from 348 speakers. Although the alignment of LB set should be provided by human in real applications when confronting a new language, in this work we simply use the forced-alignment results obtained with a full-fledged Uyghur ASR system[3].

The language model used in this work is based on trigrams trained by a text corpus involving 400k words and the lexicon involving 45000 words. The original text is written in Arabic. To simplify the processing, a simple character mapping scheme is employed to convert the Arabic characters to Latin letters.

The training and test are conducted using the Kaldi toolkit, following the THUYG20 recipe provided by CSLT[4]. More details about the recipe and the properties of Uyghur ASR systems can be found in [15].

B. Baseline

TABLE I
Baseline results

	WER%	
Model	GMM	DNN
Full UY	26.65	21.62
500 UY	55.69	72.57
CHS + Full UY	-	17.09
CHS + 500 UY	-	25.56

We first construct a couple of baseline systems. The first set of systems (Full UY) are trained using the entire Uyghur training data, by which we can estimate the performance of the system when a reasonable amount of Uyghur data are available. The second set of systems (500 UY) is trained using only the 500 labelled utterances (LB set), by which we can estimate how if the very limited data are used to train the ASR system directly. The third set of systems (CHS + Full UY) is the same as Full UY, but the DNN model is initialized from a large-scale Chinese DNN trained with 10k hours of speech; similarly, the fourth set of systems (CHS + 500 UY) is the same as 500 UY but starts from the large-scale Chinese DNN. For each set, we construct a GMM system and a DNN system trained with the criterion of cross entropy.

For the Full UY systems (with/without CHS), the number of states and pdfs are 3,376; for the 500 UY systems (with/without CHS), these numbers are 816. The less states and pdfs of the 500 UY systems are intentionally tuned to match the limited training data. The DNN model is a TDNN structure that involves 7 time-delay layers, each containing 1,200 hidden units. The context of the time delay layers is $(-20, 17)$, and the activation function is RelU.

The performance in terms of word error rate (WER) are shown in Table. I. It can be seen that using the whole training data is important to obtain a reasonable Uyghur system; using only 500 utterances can not obtain a usable system. For the Full UY systems, DNN outperforms GMM, while for the 500 UY systems, DNN is worse than GMM. Finally, using the large Chinese model as the feature extractor can significantly improve the performance for both full data condition and 500 utterance condition, demonstrating the effectiveness of transfer learning. All these observations are expected.

C. Accuracy of pseudo labels

In this section, we start to build the Uyghur monophone DNN, i.e., the Map phase of the MaR approach. A major concern here is the accuracy of the monophone DNN when producing pseudo labels for the ULB data. To evaluate the

[2]http://m2asr.cslt.org

[3]This full-fledged system is trained with the entire training data (LB+ULB). It is actually the Full UY system in Table I.

[4]https://github.com/wangdong99/kaldi/tree/master/egs/thuyg20

accuracy, we test the phone error rate (PER) of the pseudo labels, at both the frame level and the phone level. For the frame level test, the forced-alignment results produced by the fully trained DNN system (Full UY - DNN in Table I) is used as the ground truth. The initial experiment uses the monophone DNN to produce the frame-level labels frame by frame, and then merge consecutive frames with the same pseudo labels to obtain phone-level labels. The PER results are shown in Table II. It can be found that although frame-level PER is fine (21.88%), the phone-level PER is rather high (58.78%), and most of the errors are insertions. These results imply that the frame-level labels are rather noisy. This is not surprising as the pseudo label generation is independent from each frame, hence no mechanism to control spikes caused by noises and interruptions.

As discussed in Section II, a simple one-state DNN-HMM decoder can smooth the label generation, by involving both statistical constraints (HMM) and linguistic constrains (language model). This essentially utilizes the continuity of speech signals in both the acoustic and linguistic domains. The results are shown in Table II as well, where the LM used is the same as used during test, i.e., word trigrams. It shows that with the simple decoder, the phone-level PER is significant reduced (from 58.78% to 15.54%). This is a key step for the success of the MaR approach.

TABLE II
ACCURACY OF DIFFERENT LABELLING APPROACH.

	PER%	
	DNN	DNN-HMM
Frame-level	21.88	19.09
Phone-level	58.78	15.54

D. MaR results

The pseudo-labelled ULB data are used to train a GMM system and a DNN system from scratch. The network structure is the same as in the Full UY baseline system. The results are shown in Table III. For a clear comparison, the baseline results have been reproduced in the table as well. It can be observed that the MaR system, which involves both transfer learning and semi-supervised learning, obtains rather good performance, especially with the GMM framework. The WER is only slightly worse than the Full UY baseline that uses more than 50 times of labelled speech. The performance is much better than the 500 UY baseline, confirming that the semi-supervised learning has a big contribution. Another observation is that the DNN system is slightly worse than the GMM system. This might be attributed to the incorrect pseudo labels of the ULB data. These incorrect labels are supposed to impact DNN models more seriously compared to GMM models, due to the discriminative nature DNNs.

The above MaR-DNN system trains DNN model from scratch. We can also initialize the MaR DNN model using the large-scale Chinese DNN, as in the CHS + Full UY system. The performance, denoted by 'CHS + MaR' is also shown in Table III. It can be seen that the performance is significantly

TABLE III
BASELINE AND THE MAR RESULTS

	WER%	
Model	GMM	DNN
Full UY	26.65	21.62
500 UY	55.69	72.57
MaR	28.89	29.11
CHS + Full UY	-	17.09
CHS + 500 UY	-	25.56
CHS + MaR	-	22.35

improved compared to the random initialization MaR system (29.11% vs. 22.35%), conforming the contribution of transfer learning. Compared to the 500 UY transfer learning system (CHS + 500 UY), CHS + MaR is also significantly better, confirming the contribution of semi-supervised learning. Note that the pseudo labels are generated by the one-state monophone DNN system; from the results in Table III, CHS + 500 UY might be a better system for the label generation, hence more effectively using the labelled and unlabelled data. We leave this study as future work.

IV. CONCLUSIONS

We proposed a Map and Relabel approach to build speech recognition systems for languages with very few labelled utterances. A major difficulty of this almost-zero resource ASR problem is that the labelled data is too limited to train a full-fledged context-dependent system, even with transfer learning. The MaR approach solves this problem by a boosting procedure: it firstly builds a very simple one-state monophone DNN with the limited data, and then use this simple model to label untranscribed speech and train full-fledged systems with the pseudo-labelled data. This essentially combines the strength of transfer learning and semi-supervised learning. Our experiments conducted on a Uyghur database demonstrated that the MaR approach is highly effective: with only 500 utterances of speech, the MaR-GMM system can deliver a performance comparable to a GMM system trained with 50 hours of speech (28.89% vs 26.65% in WER). Unfortunately, this approach seems not suitable for DNN modeling if the pseudo-labelled data are used to train the DNN from scratch. However, when the DNN is initialized by a large-scale Chinese DNN, a reasonably good performance can be obtained.

The future work involves investigating more details of the trade-off between the contribution of labelled data (LB) and unlabelled data (ULB). Particularly, how many labelled utterances are required for a new language, and how many unlabelled data are required to obtain the performance bound. We will also investigate methods that are more suitable for DNN-based MaR, e.g., data filtering methods.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No.61633013, No.61371136.

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," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] V. Fromkin, R. Rodman, and N. Hyams, *An introduction to language*. Cengage Learning, 2018.

[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] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[5] 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.

[6] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[7] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6704–6708.

[8] Z. Tüske, J. Pinto, D. Willett, and R. Schlüter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7349–7353.

[9] K. M. Knill, M. J. Gales, A. Ragni, and S. P. Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Proc. Interspeech14*, 2014.

[10] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6975–6979.

[11] J. Gehring, Q. B. Nguyen, F. Metze, and A. Waibel, "DNN acoustic modeling with modular multi-lingual feature extraction networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 344–349.

[12] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 1225–1237.

[13] D. Wang, T. F. Zheng, Z. Tang, Y. Shi, L. Li, S. Zhang, H. Yu, G. Li, S. Xu, A. Hamdulla *et al.*, "M2ASR: Ambitions and first year progress," in *OCOCOSDA*, 2017.

[14] X. Zhu, "Semi-supervised learning literature survey," 2005.

[15] A. Rouzi, Y. Shi, Z. Zhiyong, W. Dong, A. Hamdulla, and Z. Fang, "THUYG-20: A free uyghur speech database," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 2, pp. 182–187, 2017.