Iran University of Science and Technology
Deep Learning (Spring 2019)

Transfer Learning for ASR to Deal with Low-Resource Data Problem

Maryam Asadolahzade
Department of Computer Engineering
Iran University of Science
and Technology
mar.asadolahzade@gmail.com

Abstract

End-to-end models are the state of the art for Automatic Speech Recognition (ASR) systems. Despite all their advantages, they suffer a significant problem: a huge amounts of training data are required to achieve a good performance. This is a serious challenge for low-resource languages such as Persian. Therefore, we need some methods and techniques to overcome this issue. Transfer learning is an effective method can address this problem. Our aim is to perform a phoneme recognition system for the Persian language and explore the effect of transfer learning. To this end, we first train the network with English corpus. Then, we transfer the trained network and fine-tune it with Persian corpus. Our experiments on FarsDat corpus indicate that transfer learning with a few hours of Persian data, can reduce (Phoneme Error Rate) PER by 7.88% against model trained from scratch. Moreover, this method attains improvements of 2.08% and 1.52% for PER compared with the DNN-HMM and DNN-HSMM powerful baselines, respectively.

Keywords: Automatic Speech recognition, Phoneme (Phone) recognition, Transfer learning, Phoneme Error Rate (PER), Low-resource language, Persian (Farsi) Language.

1 Introduction

Automatic Speech Recognition (ASR) task is to convert a spoken signal to the text. In this field, end-to-end approaches are the current state of the art. The important point is such approaches need huge amounts of data for training [1], [2]. For low-resource languages such as Persian, for which there is no sufficient data for training that could be an important issue. One of the technique to deal with this problem is transfer learning. This technique is applied successfully in some previous studies [1], [2]. To the best of our knowledge, transfer learning for ASR has not been applied on the Persian language, which is our goal. Therefore, we explore the effect of transfer learning for this language. Unfortunately, there is very limited research for DNN-based, in particular, end-to-end ASR systems for the Persian language. The only paper attempted to use end-to-end model for Persian is [3] which implemented a phoneme recognition system. The motivation of our work is to publish the result for end-to-end Persian phoneme recognition to alleviate future studies in this area and provide a framework for comparison for other researchers working on Persian ASR. Furthermore, we aim to find out to some extent, we can increase the current Persian ASR using transfer leaning technique. We present the evaluations of the methods using the Persian FarsDat corpus [4].

(LateX template borrowed from NeurIPS 2019)

The rest of the paper is organized as follows, Section 2 reviews previous ASR systems with transfer learning, Section 3 talks about the architecture and its input and output and the approach used in the paper. Section 4 describes datasets and performance measures used. Section 5 describes the experimental results and discussions on it. Finally, Section 6 discuss results with future research directions.

2 Related work/Background

The first attempt for transfer leaning in DNN-based ASR system was conducted in [1]. They used four European languages to build a multilingual DNN. They used one DNN such that all the hidden layers of the DNN is shared except the last softmax layer. Each language has its own softmax layer. After training with the four languages, they evaluated transfer learning with two target languages: American English and Mandarin Chinese. American English is phonetically close to the European languages used for initial network while Mandarin Chinese is different from the European languages. The results showed transferring hidden layers sharing across languages can improve accuracy for two new languages. They also concluded when the training data for target language is low, it would be better just train softmax layer instead of retrain more layers. The authors in [2] used a CNN-based end-to-end ASR. They trained the model with the English data then transfer it for the German data. They showed better performance than the German model trained from scratch. In [5], a multilingual model with 10 BABEL languages was build and it was tested for 4 other BABEL languages using transfer learning approach. Their results showed that the transfer learning from the multilingual model shows improvement over monolingual models. A more recent work [6] proposed language-adversarial transfer learning method. This method helps that the shared features contain less unnecessary language dependent information. They demonstrate promising results on IARPA Babel datasets.

3 Proposed method

Our aim is to perform a ASR system for the Persian language. To do this, we first train a ASR system for English. Then, we transfer the trained network and fine-tune it with Persian data. The overall diagram of the proposed method is shown in Figure 1.

- Input: The input of the neural network is the spectrogram extracted from the audio signal. We first segmented the signal into frames. We use 32 ms frame windows, which spanned every 8 ms. Then we applied Short-time Fourier transform (STFT) and the we mel-scaled them. We normalized features per input sequence to have zero mean and unit variance.
- Output: The output of the network is the corresponding phoneme sequence of the input audio signal.
- Architecture: We use 11 1D-convolutional layers on top of each other based on the introduced architecture in [2], [7].We use zero-padding for each layer since we aim to preserve the dimension of the input. The activation function for the first 10 layers is Relu. For the last layer, softmax activation is considered to get the probability distribution on phonemes. The last layer has 30 channel output, each of which correspond to one of the phonemes. The loss function to train network is Connectionist Temporal Classification (CTC) [8].

After training, the network predicts the probability of each phoneme for each input frame.

We did not use any beam search or decoder or even any language model in our experiments. We also use the implementation of paper [2] for our project.

4 Experimental Setup

4.1 Dataset

We use two corpora which are explained in the following:

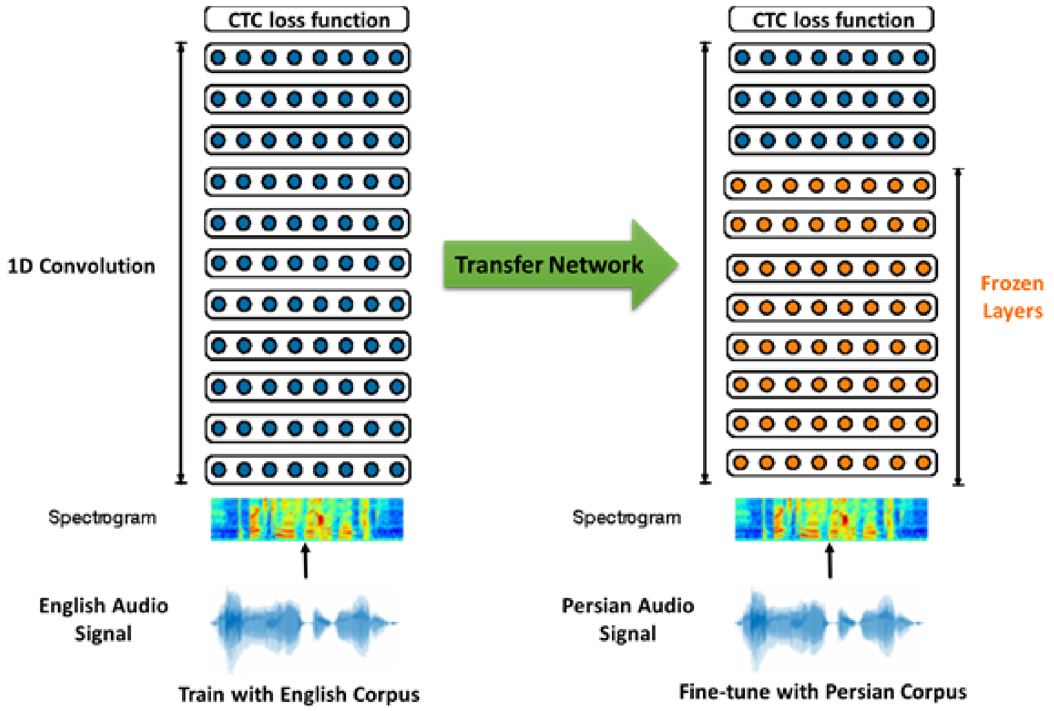

Figure 1: The diagram of the proposed method.

English corpus: For training with English language, we use the LibriSpeech corpus [9]. This corpus consists of 1000 hours of read speech, sampled at 16 kHz, from the domain of audio books. This corpus is under license ‘CC BY 4.0’ and freely available for download.

Persian corpus: We use FarsDat [4] to train the system for the Persian language, which contains the utterances of 304 female and male speakers from 10 dialect regions in Iran. Each speaker uttered 20 sentences in two sessions. This corpus is about 5 hours in size. The utterances of first 250 speakers were used as the training set and utterances of the remaining 54 speakers were used as the test set. Our experiments are speaker-independent because the speakers in the training set are different from those in the test set.

4.2 Evaluation metric

In order to evaluate a phoneme recognition system, the common measure is phoneme error rate (PER). In this measure, the recognized and reference phoneme label sequence should be compared. Two strings are compared by matching using dynamic programming. Considering the number of substitution errors, deletion errors and insertion errors shown as S, D, and I, respectively, the PER is defined as:

$$PER = \frac{S + D + I}{N}$$

where N is the total number of phonemes in the reference label [10].

5 Results

In this section, we perform three experiments to evaluate the performance. First, we investigate the effect of number of frozen layers in transferred network. Second, we evaluate the effectiveness of transfer learning by comparing the result obtained with transfer learning against ASR trained without exploiting the transferred network. Third, we compare the results with some strong/powerful baseline methods.

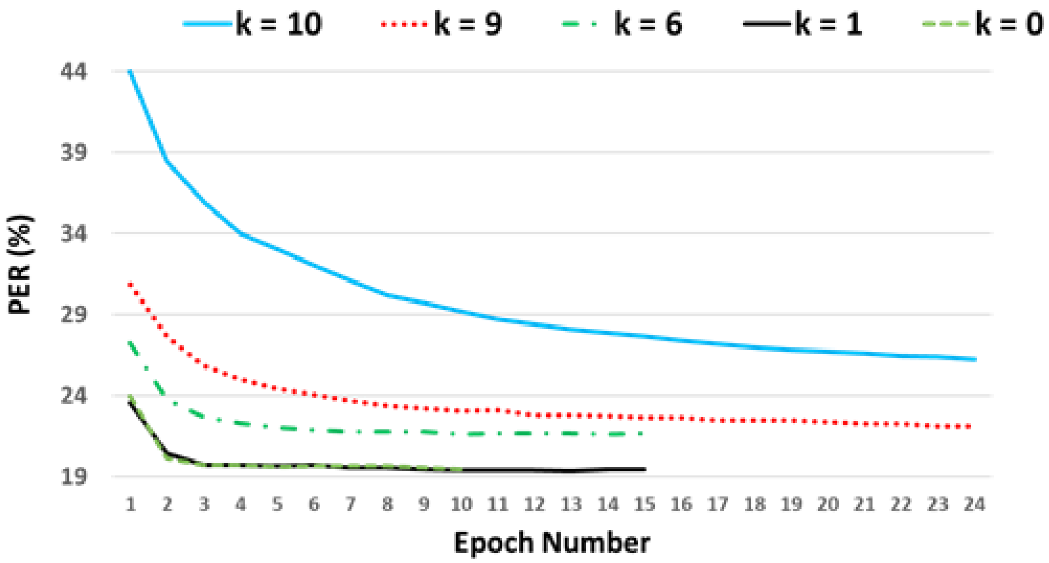

Figure 2: The performance for different values for K (number of frozen layers).

5.1 The effect of Number of frozen layers

As mentioned earlier, our network has 11 layers. To fine-tune the network with Persian corpus, we freeze K number of the lower layers, therefore the weights of 11-K top layers can update during training. To find the suitable number for K, we evaluate performance for different values for K in Figure 2.

It can be seen that for 10-frozen layers, the PER is not satisfactory. However, the difference between 10 and 9-frozen layers is really high. Therefore, when we freeze a fewer number of frozen layers, for example zero or one, we can get the better result. The reason is that English and Persian languages have some audio properties in common but it is necessary to fin-tune all the layers to reach the best results. Our results are consistent with finding in [2].

5.2 The effect of transfer learning

In this experiment, we compare the results of transfer learning with the model without transferring, i.e. we train the network with just Persian corpus. We demonstrate the effect of transferring in terms of both training time and accuracy in Figure 3.

The results show training from scratch convergence slower than all of the transferred network, even 10-frozen layer transfer network. Besides the accuracy is lower than transferred network. In order to compare the results precisely, we report them in Figure 4. The results are based on 10 epochs of training.

As expected, transfer learning outperforms model trained from scratch even with few number of frozen layers.

5.3 Comparing with baselines

To have a better comparison with other methods, we compare our end-to-end model with some baselines including Gaussian mixture model-hidden Markov (GMM-HMM), deep neural network-hidden Markov model (DNN-HMM) and deep neural network-hidden semi-Markov model (DNN-HSMM), which are explained in detail in our previous work [11].

From the results, we can see that end-to-end system without transferring yields worse performance than baselines even though it is the state of the art. Because the size of Persian corpus is extremely low. However, we observe significant improvement for the system trained by transfer learning, which is expected.

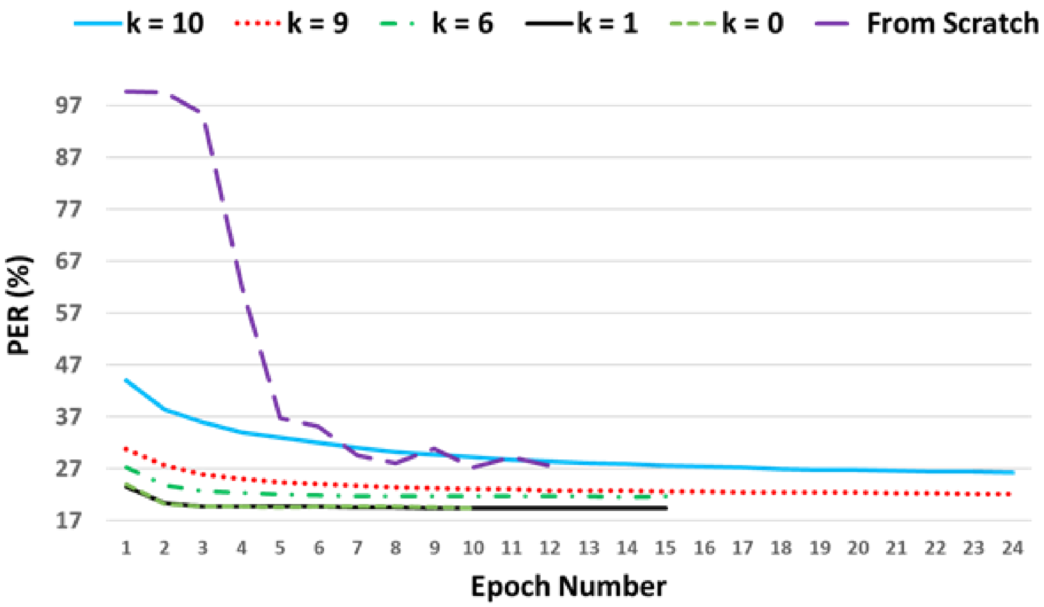

Figure 3: Comparing the results of transfer learning and learning from scratch.

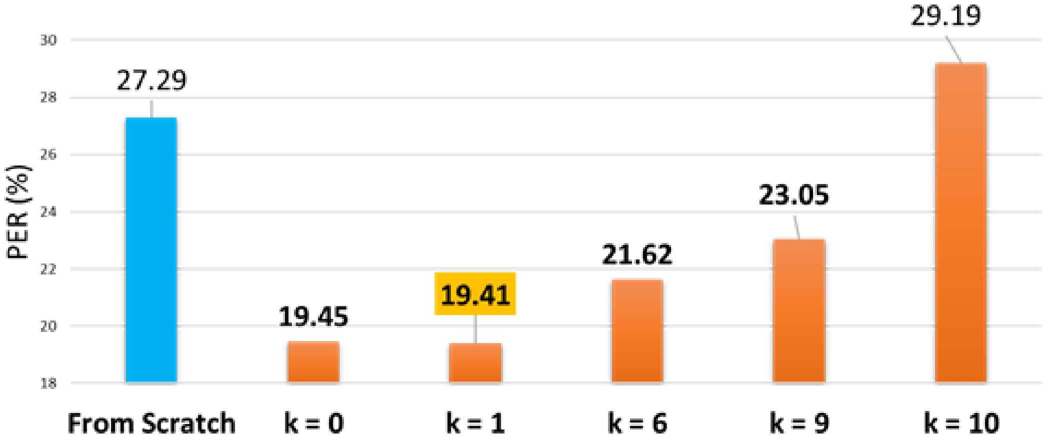

Figure 4: Comparing the results of transfer learning and learning from scratch.

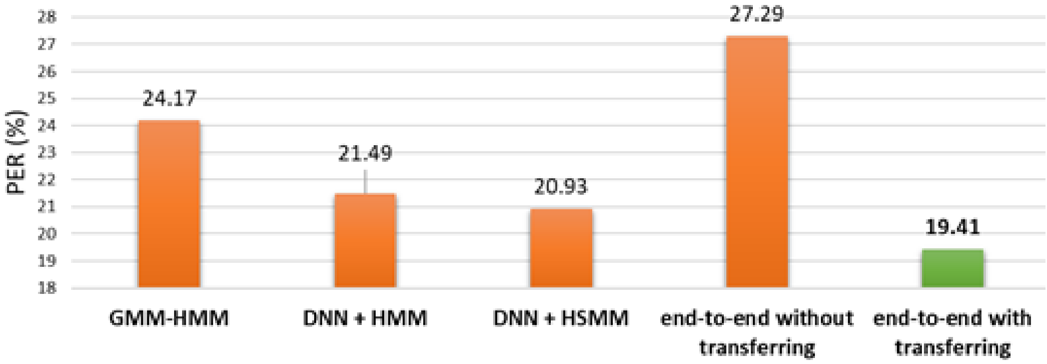

Figure 5: Comparing the results of transfer learning with some baselines.

6 Discussion

We could obtain three benefits in this paper. Firstly, we provide a comparative framework and an end-to-end model for future studies in Persian ASR. Secondly, we reduce training time using transfer learning. Transfer learning can be seen as a kind of weight initialization. Thus, instead of training network with random weights, with the help of transfer learning we can start learning from suitable values for weights which are trained previously using a huge corpus. Lastly, we obtain better results compared to model trained from scratch and baseline methods.

For future work, we will use language model during decoding to achieve additional improvement may.

References

[1] 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 ICASSP*, 2013, pp. 7304–7308.

[2] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," *2nd ACL Workshop on Representation Learning for NLP*, 2017.

[3] S. Alisamir, S. M. Ahadi, and S. Seyedin, "An end-to-end deep learning model to recognize Farsi speech from raw input," *in 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2018, pp. 1–5.

[4] M. Bijankhan, J. Sheikhzadegan, and M. R. Roohani, "FARSDAT-The speech database of Farsi spoken language," *presented at the The Proceedings of the Australian Conference on Speech Science and Technology*, 1994.

[5] J. Cho et al., "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," *in Spoken Language Technology Workshop (SLT)*, 2018, pp. 521–527.

[6] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Language-adversarial transfer learning for low-resource speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 3, pp. 621–630, 2019.

[7] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System," *arXiv:1609.03193* , Sep. 2016.

[8] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," *in Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, pp. 369–376.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," *in ICASSP*, 2015, pp. 5206–5210.

[10] S. Young et al., *The HTK book*, vol. 3. 2002.

[11] M. Asadolahzade Kermanshahi and M. M. Homayounpour, “Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM,” *Journal of AI and Data Mining*, vol. 7, no. 1, pp. 137–147, 2019.