INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Transfer Learning and Distillation Techniques to Improve the Acoustic Modeling of Low Resource Languages

Basil Abraham, Tejaswi Seeram, S. Umesh

Indian Institute of Technology-Madras, India
{ee11d032,ee15s044,umeshs}@ee.iitm.ac.in

Abstract

Deep neural networks (DNN) require large amount of training data to build robust acoustic models for speech recognition tasks. Our work is intended in improving the low-resource language acoustic model to reach a performance comparable to that of a high-resource scenario with the help of data/model parameters from other high-resource languages. we explore transfer learning and distillation methods, where a complex high resource model guides or supervises the training of low resource model. The techniques include (i) multi-lingual framework of borrowing data from high-resource language while training the low-resource acoustic model. The KL divergence based constraints are added to make the model biased towards low-resource language, (ii) distilling knowledge from the complex high-resource model to improve the low-resource acoustic model. The experiments were performed on three Indian languages namely Hindi, Tamil and Kannada. All the techniques gave improved performance and the multi-lingual framework with KL divergence regularization giving the best results. In all the three languages a performance close to or better than high-resource scenario was obtained.

Index Terms: speech recognition, low-resource languages, transfer learning, distillation

1. Introduction

Deep neural networks (DNN) have become quite ubiquitous in various deep learning applications including automatic speech recognition (ASR) tasks in the recent past [1]. An important facet owing to these improvements, as seen even in commercial voice assistants like Siri, Alexa etc, is the availability of huge amount of resources for training the DNNs. This makes it possible to estimate the millions of parameters without overfitting issues. Google voice [2], Baidu etc use thousands of hours of transcribed speech data for building acoustic modules. Evidently, robust speech recognition systems in the world are limited to popular major languages like English, French etc. The robustness of a system is hugely dependent on the availability of sufficiently large amount of transcribed data. Data sparsity is one of most challenging problems in building DNN based acoustic models. Acquiring data along with their proper transcription is costly in terms of man hours. Languages with adequate amount of training data are referred to as high resource languages and those with limited amount of data are low resource languages [3].

Simpler neural networks can be built using the sparse data without overfitting, however, promising performances are observed with deeper and wider networks. Prominent approaches this issue include sharing of data and model parameters from high resource data sets. Frameworks like multi-task learning [4, 5], transfer learning [6, 7, 8, 9, 10] etc were made to handle scarcity of data availability in different frameworks showing visible improvements in performance. In the transfer learning paradigm, we use the multi-lingual DNN training with separate softmax for each language at the output layer of DNN. We refer to this method as *blocksoftmax* [11]. Subspace Gaussian mixture models [12] are one of the prominent methods of sharing model parameters. Multilingual modeling techniques involving training of common models for high and low resource languages have also shown relatively improved performance. Nonetheless, these sharing of information requires overlapping of phone sets, else, the network might get confused (more than learn) with varying context dependencies across languages.

In this paper we propose a technique to overcome the data sparsity problem in building DNN based acoustic models. We apply this in transfer learning and distillation frameworks. Kullback-Leibler (KL) divergence based criterion is used in terms of the high-resource acoustic model is added as an additional constraint in blocksoftmax framework for low-resource acoustic modeling. The KL divergence based constraint avoids the the issue of overfitting by acting as a regularization term in the loss function during the multilingual training of blocksoftmax with low-resource and high-resource language. In this paper we also use the generalized distillation framework [1, 13, 14, 15] where the complex high-resource model can guide the training of low-resource language model.

The techniques were applied to three under-resourced Indian languages, Hindi, Tamil and Kannada. In each language we considered two low-resource data set 10 and 5 hours, respectively and a high-resource dataset consisting of 50 hours of training data. Experiments were performed with one of the languages as a low-resource and the other two as high-resource. In all the three languages we obtained consistent improvements in all the three techniques. Among the proposed multilingual techniques with KLD and distillation, the KLD based criterion gave better performance.

The paper is organized as follows. Section 2 describes the proposed technique in the case low resource languages and the corresponding frameworks investigated on. Section 3 gives the experimental set-up and the corresponding results. Section 4 gives the analysis and section 5 concludes the paper.

2. Proposed Technique

In this paper we propose techniques to include additional criterion in the loss function while training the DNN acoustic model of a low-resource language. This criterion involves giving additional constraint to update the parameters using the posterior distribution of a DNN trained over high resource language data. This constraint moves the weight parameters of the DNN model conservatively towards the target language, helping with the issues of overfitting by acting as a regularization term over the distribution during the gradients back-propagation. This additional cross-entropy term coming from the high resource model

http://dx.doi.org/10.21437/Interspeech.2017-1009

acts as a KL (KullbackLeibler) divergence based measure (ignoring the terms independent of parameters) between the distributions of the low-resource and high-resource acoustic models.

$$L = \lambda \sum_{t=1}^{T} \sum_{i=1}^{N} y_i \log p(s_i/x_t) + (1-\lambda) \sum_{t=1}^{T} \sum_{i=1}^{M} \hat{y}_i \log p(s_i/x_t)$$

where L is the loss funtion (weighted cros entropy loss functions), λ is weight parameter, s_i is a senone, y_i is a senone label (hard label) for an input acoustic observation x_t , $\hat{y}$ corresponds to the posterior distribution obttained from a high resource DNN and $p(s_i/x_t)$ is the estimated posterior probability for the low resource data. T, N and M are the total number of observations, number of senones in low and high-resource languages, respectively.

Unlike in the conventional techniques of using KL divergence based constraints like in speaker adaptation [16], its use is limited in the multilingual scenario owing to differences in context-dependent states in both the languages. Hence, in the proposed technique we overcome this limitation by introducing an additional layer before the softmax function specific to each language.

2.1. Multilingual training with KLD

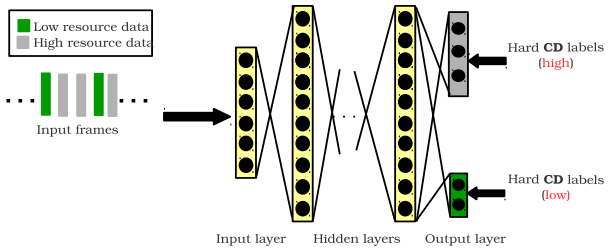

Figure 1: *Block-softmax model*

The proposed technique was applied in the multilingual framework commonly known as blocksoftmax modeling. The block schematic of the blocksoftmax acoustic modeling technique is given in Figure 1. In this technique the training data from both the high and low-resource languages are given at the same time with separate softmax activations for each language at the output layer. The cross-entropy errors for each language data are generated only by the corresponding softmax layer, hence the name blocksoftmax model. The block schematic of the proposed approach is given in Figure 2. In the proposed technique we add an additional constraint to the blocksoftmax model given in Figure 1. The additional KL divergence based constraint is added as another output target from high-resource model probability distribution in terms of the low-resource data. These targets are obtained by forward passing the low-resource data through the high-resource DNN model. We call these targets soft-labels, since each has more than one non-zero values unlike the conventional hard-targets (one-hot vectors). Also, intuitively, since the the deviation of the posterior distribution of the high resource model is closer to its true distribution compared to that of the low resource model, this additional cross-entropy term forces the low resource block output to not deviate too much from its true distribution.

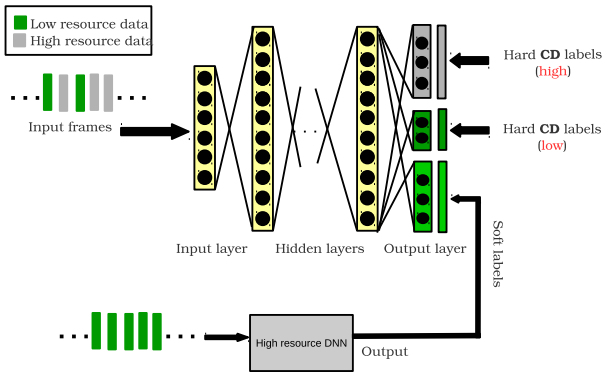

Figure 2: *Adding KL divergence to block-softmax*

$$\hat{L} = \lambda \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{y}_i \log p_1(s_i/x_t) + (1-\lambda) \sum_{t=1}^{T} \sum_{i=1}^{M} z_i \log p_2(s_i/x_t) \quad (1)$$

where y_i, z_i represent the hard labels corresponding to the low and high resource languages, respectively. $\tilde{y}_i = \eta y_i + (1-\eta) p_1^H$, p_1^H is the forward pass output of the low resource data using the high resource model. p_2 corresponds to the high resource posterior estimate.

In this framework, the loss function is a weighted sum of negative cross entropy errors for each language at the individual softmax block. The errors are back-propagated from each block according to the weights assigned. The method combats the performance gap occurring in the DNN model when trained only with the low resource language separately. The hidden layers are shared across the languages and the network, therefore, has ample data additional to the low resource data to update the parameters.

2.2. Generalized Distillation

Generalized Distillation [1, 13, 14, 15] is another method where a constraint from high-resource model can be added in training an acoustic model for a low-resource language. The set-up is given in Figure 3. In distillation framework an intelligent teacher provides *easier* targets for the student model by providing information about regularities in the data in the form of target posteriors.

In distillation framework the blocksoftmax model with low-resource softmax block alone was used as the teacher. The block schematic of the proposed setup is given in Figure 3. The teacher model was used to generate the soft-labels by passing the low-resource data through it. In this case we have not used the temperature parameter to soften the forward passed output.

3. Experimental Setup

In this work, the experiments were performed using three Indian languages, namely, Tamil, Hindi and Kannada from MANDI database [17, 10]. Experiments were performed with every combination of languages, as in, one language at a time is of low

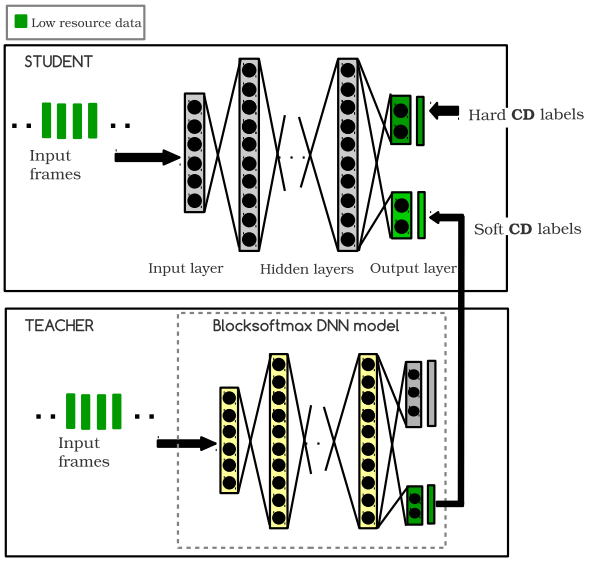

Figure 3: *Distillation framework*

resource and the other two are of high-resource. This was done to study the portability of data and model parameters for each low-resource language. Kaldi toolkit [18] was used to build the acoustic models in all the cases.

3.1. Database

The experiments in this paper were performed in three low-resource Indian languages, Hindi, Tamil and Kannada from the MANDI database. MANDI database is a multi-lingual database consisting of 12 Indian languages collected for "Speech based access to agricultural commodity prices", a Government of India project to build ASR systems in Indian languages to provide the information of different agricultural commodities or weather information in a particular place in India. The speech data was collected from the farmers in their work place. Hence, the speech data has varied environmental conditions from quiet to very noisy. All the multilingual experiments were carried out using global phone-set containing the phones of all the languages and therefore, could be shared across the languages [19]. The lexicon used ILSL-12 format for phone mapping [17].

Among the 12 languages, we considered three languages Hindi an Indo-Aryan Central origin, while Tamil and Kannada are of Dravidian origin. For each language we made two low-resource data sets consisting of 10 and 5 hours respectively and the high-resource data set consisting of 50 hours of training data. The low-resource data sets are randomly obtained from the high-resource data. The Hindi and Kannada database consist of short phrases and Tamil database has some short sentences. From observation, Tamil database was found to be more noisy compared to other two.

3.2. Baseline Models

All the baseline results for each language in the low resource set-up are given in the corresponding tables. The DNN models were trained with 6 layers and 2048 nodes per layer (owing to the limited vocabulary size). The context dependent state alignments were obtained using CDHMM (continuous density hidden Markov model) models trained on mel frequency cepstral coefficients (MFCC) extracted from the respective data using a frame length of 25ms and frame shift of 10ms, augmented with first and second order derivatives. The DNN models used log-filter bank ($40 + \Delta + \Delta\Delta$) features for training. Input features were stacked over a context window of 11 frames (± 5 splicing). In an initial unsupervised stage of training, layer-wise pretraining was done deep belief network (DBN)(stack of restricted Boltzmann machine) using the entire data.

3.3. Experiments with Multi-lingual modeling with KLD

As described earlier, experiments were done in blocksoftmax framework with additional divergence constraint over the output distribution of the low resource softmax activation. In blocksoftmax set-up, one low resource language at a time is trained along with one or more other high-resource languages. Experiments were performed for all the languages in both $5hr$ and $10hr$ cases for each. 40-dimensional log-filterbank features extracted from the input frames were used as input to the network with delta and acceleration coefficients. The network was trained with 6 layers with 2048 nodes in each hidden layer. A separate output layer was used for each language with dimension equal to the context-dependent states in each language (the state alignments are given from the CDHMM model trained using the corresponding data). The network was jointly optimized with data from both languages as described in Section 2.1.

The proposed KLD experiments were also performed with the similar network structure at the hidden layers. However, the difference comes in the output layer where an additional constraint is given as targets with soft-labels generated from a well trained high-resource model. In this case also the network is jointly optimized with both languages as described in Section 2.1. These soft targets are basically posterior distribution of the high resource model when the low resource input is fed to it.

3.4. Experiments with Generalized Distillation

Generalized distillation experiments were performed in a multilingual framework as described in Section 2.2. The experiments were performed with each language as low-resource language. In every case, a blocksoftmax model trained with the corresponding low-resource language with other high-resource language(s) is used as the teacher model. The student model is a DNN to be trained using the low resource language. In order to initialize the student model better, hidden layer parameters of a high resource model are borrowed. The output layer of this initialized network is trained using input frames from low resource language with targets (additional to its context dependent (CD) hard labels) coming from the teacher model. This posterior distribution correspond to soft labels and is expected to provide the student model some privilege information (distilling knowledge) about the phoneme structures obtained from the high resource model.

3.5. Experimental Results

The recognition performance for the multilingual model with KLD are given in Tables 1, 2 and 3. The experiments were performed under varying amount of low-resource data. The experiments were performed in all the three languages with Train-10hr and Train-5hr as low-resource language data sets. In both the case the proposed technique of adding KLD to the blocksoftmax is giving improved recognition performance over the blocksoftmax model.

The recognition performance of the distillation experiments in multilingual framework are given in Tables 1, 2 and 3. The

Table 1: *Recognition Performance of Hindi Low-resource Data*

Training Data	Hindi_10hr			Hindi_5hr	
	Block-softmax		Distill-ation	Block-softmax	
	no KLD	KLD		no KLD	KLD
Hindi_50hr*	9.30	-	-	-	-
Hindi_50hr* (Hindi_10hr tree)	12.00	-	-	-	-
Hindi_low*	16.14	-	-	21.9	-
Tamil_50hr + Hindi _low	14.31	13.79	14.19	18.49	17.75
Kannada_50hr + Hindi_low	12.83	12.69	12.90	16.78	**16.69**
Tamil_50hr + Kannada_50hr + Hindi_low	**12.05**	12.5	12.30	18.15	17.48

* Baseline monolingual DNN model

Table 2: *Recognition Performance of Tamil Low-resource Data*

Training Data	Tamil-10hr			Tamil-5hr	
	Block-softmax		Distill-ation	Block-softmax	
	no KLD	KLD		no KLD	KLD
Tamil_50hr*	11.98		-	-	-
Tamil_50hr* (Tamil_10hr tree)	14.12		-	-	-
Tamil_10*	15.76		-	18.35	-
Hindi_50 + Tamil_10	13.54	13.90	13.30	15.38	14.74
Kannada_50 + Tamil_10	13.01	12.77	12.90	14.99	14.82
Hindi_50 + Kannada_50 + Tamil_10	12.49	**12.40**	12.45	14.40	**14.37**

* Baseline monolingual DNN model

Table 3: *Recognition Performance of Kannada Low-resource Data set*

Training Data	Kannada_10hr			Kannada_5hr	
	Block-softmax		Disti-llation	Block-softmax	
	no KLD	KLD		no KLD	KLD
Kannada_50*	5.41		-	-	-
Kannada_50* (Kannada_10hr tree)	5.78		-	-	-
Kannada_10*	7.41		-	10.83	-
Tamil_50 + Kannada_10	6.94	6.85	6.85	8.90	8.79
Hindi_50 + Kannada_10	6.57	6.49	6.56	8.43	8.31
Tamil_50 + Hindi_50 + Kannada_10	6.47	**6.35**	6.40	7.97	**7.84**

* Baseline monolingual DNN model

Table 4: *Consolidated Recognition Results (WER%)*

Train-set	Base-line	Best Model	RI (%)	Block-softmax	KLD
Hindi-10hr	16.14	12.05	25.34	✓	
Tamil-10hr	15.76	12.40	21.32		✓
Kannada-10hr	7.41	6.35	14.30		✓
Hindi-5hr	21.90	16.69	23.79		✓
Tamil-5hr	18.35	14.37	21.69		✓
Kannada-5hr	10.83	7.84	27.6		✓

* RI: Relative improvement

performance obtained is inferior to that of the proposed technique of blocksoftmax using KLD. The distillation experiments were also done with both Train-10hr and Train-5hr data sets in all the three languages. Since in both the cases, the distillation results were inferior to block-softmax with KLD. *Hence, only the Train-10hr results are given in the tables mentioned due to space constraints.*

4. Analysis

From all the experiments mentioned in Sections 3.3 and 3.4, we analyze the following:

- KLD upon multilingual framework works the best compared to the conventional techniques of sharing data/model parameters like blocksoftmax and generalized distillation.
- In all the three languages, a considerable relative improvement is observed over their corresponding monolingual DNNs.
- With all the techniques described, we achieve a performance very close to their corresponding high-resource counterparts in case of Train-10hrs. The Table 4 shows the best recognition performance obtained in all cases.
- In all cases, using both the high-resource languages benefited the low-resource language the most.
- The results are in par with the similarity among languages that the native speakers observe. The low-resource Hindi is benefited more by Kannada compared to Tamil. Similarly, low-resource Tamil from Kannada and low-resource Kannada from Hindi benefit the most.

5. Conclusion

In this work, we address the issue of inadequate transcribed training data availability for building robust DNN acoustic models in Indian languages. With the proposed technique of using additional KL divergence based constraint in blocksoftmax framework, an improved recognition performance is observed. A performance close to high-resource DNN module is achieved using the proposed technique for the low-resource cases. On an average, a relative improvement of above 20% is seen.

6. Acknowledgements

This work was supported in part by the consortium project titled "Speech-based access to commodity price in six Indian languages", funded by the TDIL program of DeITY of Govt. of India. The authors would like to thank consortium members involved in collecting Hindi Tamil, and Kannada corpus.

7. References

[1] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.

[2] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Your word is my command: Google search by voice: A case study," *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Chapter 4, Springer*, pp. 61–90, 2010.

[3] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[4] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[5] V. B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to vietnamese language," *IEEE Trans. Audio, Speech & Language Processing*, vol. 17, no. 8, pp. 1471–1482, 2009.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010.

[7] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2015, Hong Kong, December 16-19, 2015*, 2015, pp. 1225–1237.

[8] 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*. IEEE, 2013, pp. 7304–7308.

[9] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *Proc. ICASSP*, 2010, pp. 4334–4337.

[10] B. Abraham, S. Umesh, and N. M. Joy, "Overcoming Data Sparsity in Acoustic Modeling of Low-Resource Language by Borrowing Data and Model Parameters from High-Resource Languages," in *Proc. INTERSPEECH*, 2016.

[11] L. M. Marcos and F. Richardson, "Multi-lingual deep neural networks for language recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 330–334.

[12] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The Subspace Gaussian Mixture Model - A Structured Model for Speech Recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[13] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5-6, pp. 544–557, 2009.

[14] K. Markov and T. Matsui, "Robust speech recognition using generalized distillation framework," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 2364–2368.

[15] V. Vapnik and R. Izmailov, "Learning using privileged information: similarity control and knowledge transfer," *Journal of Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[16] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP 2013*, 2013.

[17] Basil Abraham, Neethu Mariam Joy, Navneeth K., and S. Umesh, "A Data-Driven Phoneme Mapping Technique Using Interpolation Vectors of Phone-Cluster Adaptive Training," in *Proc. SLT*, December 2014, pp. 36–41.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, December 2011.

[19] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. Int. Conf. Spoken Language Processing (ICSLP*, 1998, pp. 1819–1822.