Few-shot learning for cross-lingual end-to-end speech recognition

Dhanya Eledath[1], *V. Pavithra*[2], *Narasimha Rao Thurlapati*[2], *Tirthankar Banerjee*[1],
V. Ramasubramanian[1]

[1]International Institute of Information Technology - Bangalore (IIIT-B)
[2]Samsung R&D Institute, Bangalore (SRIB)

dhanya.eledath@iiitb.ac.in

Abstract

This paper explores cross-lingual end-to-end (E2E) continuous speech recognition in a few-shot learning framework. 'Few shot learning' (FSL) is an emerging area of research which imitates the human-way of learning new concepts from few examples leveraging on previous experience. We adapt a model based prior-knowledge FSL paradigm called 'Matching Networks' (MN) for the task of continuous speech recognition by integrating MN architecture with CTC-loss based end-to-end training and decoding. The robustness of FSL-MN framework lies in a cross-lingual training-inference scenario, where efficient embedding functions are learnt from a large training corpus in one domain and such learnt embedding functions are used as prior knowledge to perform few shot inference on a small target data. We utilize this aspect of MN in developing an E2E cross-lingual phoneme/grapheme recognition system for low resource languages. We validate the cross-lingual MN advantage using two sets of experiments: 1) high resource Hindi speech corpus (50 hrs) used in the continuous speech decoding of low resource Indian languages - Gujarati and Marathi, 2) Tamil (104 hrs) source model on the cross-lingual inference of Malayalam and Kananda. Our proposed architecture consistently exhibits a superior performance over a baseline Bi-LSTM transfer learning setting for cross-lingual scenarios.

Index Terms: few-shot learning, matching networks, continuous speech recognition, cross-lingual end-to-end ASR

1. Introduction

End-to-end automatic speech recognition (E2E-ASR) systems have evolved dramatically with the emergence of deep learning based sequence to sequence networks. E2E architectures employ large number of network parameters and hence require abundant supervised training data to leverage the underlying optimization aspects. But not all languages have such large training corpora available and one of the challenges is to devise methods to easily adapt existing ASR systems to new languages with minimal effort at reasonable cost. Specifically, we address the task of dealing with languages having few hours of transcribed speech data available and which fall under the category of 'low resource languages'. Limited data availability pose difficulties in developing efficient speech recognition systems for low resource languages. 'Transfer learning' (TL) methods and 'multilingual' training using pooled data and common phoneset from different languages are the most commonly used approaches for improving the performance of low-resource speech recognition systems [1]. In this paper, we adapt a new machine learning paradigm 'Few-Shot Learning' (FSL) [2] to develop low-resource continuous speech recognition systems.

Learning efficiently from few samples [3] or few-shot learning (FSL) is emerging in a radical direction in machine learning with a variety of paradigms and network realizations. FSL has gained great momentum in the areas of image classification [4], image retrieval [5], gesture recognition [6], language modeling [4]. The FSL methods belong to the class of meta learning frameworks wherein the prior knowledge acquired from distinct but similar tasks is used to learn a new task quickly using very few shots per class. All current FSL approaches use prior knowledge of various kind (e.g. data, model and algorithm) to reduce the so-called 'sample complexity' defined as the number of training samples needed to guarantee minimizing empirical risk [2].

Recently, in a first of its kind attempt, we applied few-shot learning for frame-wise phoneme recognition by adapting a 'model-based' FSL paradigm 'matching networks' (MN) [7]. In the present work, we extend this MN framework to end-to-end continuous speech recognition in a formulation termed MN-CTC network. This uses a Connectionist Temporal Classification (CTC) [8], [9] loss based end-to-end (E2E) training of matching networks (MN) and an associated CTC-based decoding of continuous speech. As a primary contribution here, we examine and adapt this MN-CTC framework within a cross-lingual setting. This is in line with the cross-domain applicability of MN's theory, now adapted for E2E continuous speech recognition under 'low' resource conditions. Here, the network is trained in a possibly high-resource language, and applied, in a cross-domain manner, for decoding continuous speech of a 'low' resource target language, with few examples of input acoustic frames/class label (characters in the target language transcripts).

Keeping in view that matching networks for FSL can be construed as transfer learning, we note there have indeed been traditional transfer learning frameworks for acoustic-model adaptation and learning for a target language with low resources usually in a multi-lingual setting or cross-lingual training approaches. In such work, e.g. [10], [11], [12], [13], [14], [15], [16], [17] the hidden layers are either HMM or DNN models which are shared by multiple languages, and the output layer is either language specific or universal IPA-based phone set. Usually, the hidden layers are transferred from a network trained on a high resource language, on to a low resource training condition, which involves adding the discriminative output layer with the target language phone set and fine-tuning of the entire network weights with the target low resource training data. In contrast, the proposed MN-CTC involves learning of embedding functions from a training support set which has class labels which are not present in the test set without further setting up the target phone-set in a loss function and in principle without re-training on the test support set. Another FSL related work in speech recognition is a model-agnostic meta-learning algorithm (MAML) based MetaASR [18] designed for multilingual speech recognition using sequence-to-sequence architecture.

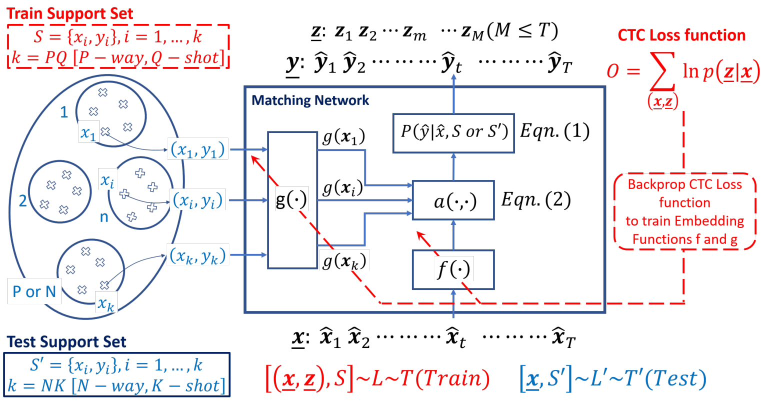
Train Support Set
$S = \{x_i, y_i\}, i = 1, \ldots, k$
$k = PQ\ [P - way, Q - shot]$
1
2
n
P or N
x_1
x_i
x_k
(x_1, y_1)
(x_i, y_i)
(x_k, y_k)
Test Support Set
$S' = \{x_i, y_i\}, i = 1, \ldots, k$
$k = NK\ [N - way, K - shot]$
Matching Network
$g(\cdot)$
$g(\boldsymbol{x}_1)$
$g(\boldsymbol{x}_i)$
$g(\boldsymbol{x}_k)$
$P(\hat{y}|\hat{x}, S\ or\ S')$
Eqn. (1)
$a(\cdot,\cdot)$
Eqn. (2)
$f(\cdot)$
$\underline{\boldsymbol{z}}$: $\boldsymbol{z}_1\ \boldsymbol{z}_2 \cdots \boldsymbol{z}_m\ \cdots \boldsymbol{z}_M (M \leq T)$
$\underline{\boldsymbol{y}}$: $\hat{\boldsymbol{y}}_1\ \hat{\boldsymbol{y}}_2 \cdots\cdots\cdots \hat{\boldsymbol{y}}_t\ \cdots\cdots\cdots \hat{\boldsymbol{y}}_T$
$\underline{\boldsymbol{x}}$: $\hat{\boldsymbol{x}}_1\ \hat{\boldsymbol{x}}_2 \cdots\cdots\cdots \hat{\boldsymbol{x}}_t\ \cdots\cdots\cdots \hat{\boldsymbol{x}}_T$
CTC Loss function
$O = \sum_{(\underline{x},\underline{z})} \ln p(\underline{z}|\underline{x})$
Backprop CTC Loss function to train Embedding Functions f and g
$[(\underline{x}, \underline{z}), S] \sim L \sim T(Train)$
$[\underline{x}, S'] \sim L' \sim T'(Test)$

Figure 1: *MN-CTC network: Matching Networks with CTC loss for End-to-end Continuous Speech Recognition*

2. Matching Networks (MN)

Matching networks (MN) [4] belongs to the class of metric-learning based FSL approaches whose central objective is to learn a similarity metric between the few-shot labelled samples and test sample in an embedded space. This similarity metric in MN takes the form of a cosine-similarity based attention mechanism. MN employs an episodic training strategy to learn embedding functions, parameterized using deep neural networks, from training samples drawn from classes that are not part of the test problem, i.e., the test class belongs to a set of classes not seen during the training of the embedding functions.

MN framework is composed of two parts, the training part which learns the embedding functions and the inference part that uses the learnt embeddings as prior knowledge to classify a new test sample. The core idea of MN [4] was to perform a N-way K-shot classification, i.e to classify a test sample $\hat{\mathbf{x}}$ (e.g. an image) as one of N class labels (e.g. visual objects), by using K-shot examples.

2.1. Matching networks notations

Every few-shot classification task in MN formulation uses a support set consisting of few labeled examples per class called the support-set and a query set comprising of test samples each of which is to be classified as one among the support set samples. The MN formulation dictates that the support-set be sampled both in training and inference for each batch-query and inference-query. This is to ensure episodic training and matched conditions between training and inference. In our work, we refer to the inference (test) support set as S' formed from N classes with K samples each, i.e., $S' = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{k}$ where $k = NK$, $\mathbf{x}_i$ is the test support sample and $\mathbf{y}_i$ is the one-hot encoded representation of class labels. The test sample $\hat{\mathbf{x}}$ is mapped into a probability distribution over output labels $\hat{y}$ by performing the KDE/KNN generalization posterior estimate as in Eqn. 1,

$$\hat{\mathbf{y}} = P(\hat{y}|\hat{\mathbf{x}}, S') = \sum_{i=1}^{k} a(\hat{\mathbf{x}}, \mathbf{x}_i)\mathbf{y}_i \tag{1}$$

Here, the posterior vector $\hat{\mathbf{y}}$ or equivalently $P(\hat{y}|\hat{\mathbf{x}}, S')$ - the posterior probability distribution over the set $\hat{y}$ (the set of N class labels the test sample can belong to) is estimated as a linear combination of the labels in the test support set S', with the linear combination weights as the attention weights a, which is essentially a metric between $\hat{\mathbf{x}}$ and each of the few-shot samples $\mathbf{x}_i$ in S'. A maximum a posterior (MAP) prediction of the posterior vector $\hat{\mathbf{y}}$ yields the class label to which $\hat{\mathbf{x}}$ is classified. The $a(.,.)$ is defined by Eqn. 2 as the softmax over the cosine similarity c with embedding functions f and g being appropriate neural networks (possibly with $f \neq g$) to embed $\hat{\mathbf{x}}$ and $\mathbf{x}_i$ respectively.

$$a(\hat{\mathbf{x}}, \mathbf{x}_i) = \frac{e^{(c(f(\hat{\mathbf{x}}), g(\mathbf{x}_i)))}}{\sum_{j=1}^{k} e^{(c(f(\hat{\mathbf{x}}), g(\mathbf{x}_j)))}} \tag{2}$$

3. Adaptation of Matching Networks to E2E continuous speech recognition

Here, we first present the adaptation of the MN formulation to end-to-end (E2E) continuous speech recognition (CSR) using CTC loss in a network termed MN-CTC proposed by us as shown in Figure 1 for both training of the MN-CTC network and decoding of input continuous speech into phoneme and grapheme sequence on which PER and LER is computed respectively as a performance metric. The adaptation of the MN framework to CSR in the MN-CTC network for both training and decoding are as below. The adaptation of the MN-CTC to the cross-lingual scenario which forms the main contribution of this paper is given in Section 4.

3.1. MN-CTC Training

The MN training learns the embedding functions f and g from a ‘training’ support set $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{k}$ defined as a P-way, Q-shot set, i.e., with $k = PQ$. S comprises P classes not seen in the test-support S' (as defined in Section 2.1). Q is the number of examples per class in S. The support set S is generated first from the posteriors in a BiLSTM-CTC decoding of the continuous speech utterances in the training set $S \sim T$ - reconstituted into a P-way, Q-shot support set S. By this, one of the P classes of S is the ‘blank’ symbol ‘_’ and $\mathbf{y}_i$ for any $\mathbf{x}_i \in S$ is a one-hot encoding vector with one of the components standing for the ‘_’ label. The MN-CTC network is trained end-to-end using the CTC loss function $\sum_{(\underline{\mathbf{x}}, \underline{\mathbf{z}}) \in B} -\log P_\theta(\underline{\mathbf{z}}|\underline{\mathbf{x}})$, where $\underline{\mathbf{z}}$ is the paired phoneme/grapheme-label sequence ground truth of the input continuous speech feature vector sequence $\underline{\mathbf{x}}$: $\hat{\mathbf{x}}_1$, $\hat{\mathbf{x}}_2$, ..., $\hat{\mathbf{x}}_t$, ..., $\hat{\mathbf{x}}_T$ in batch B sampled from the training data T. The MN training finds the optimal network parameter $\theta = (f, g)$ through back-propagation (shown in red-dashed lines in Figure 1) by minimizing the objective function as in Eqn. 3. Note that $(\underline{\mathbf{x}}, \underline{\mathbf{z}}) \notin S$

$$\theta = \arg\min_{\theta} E_{L \sim T}\left[E_{S \sim L, B \sim L}\left[\sum_{(\underline{\mathbf{x}}, \underline{\mathbf{z}}) \in B} -\log P_\theta(\underline{\mathbf{z}}|\underline{\mathbf{x}})\right]\right] \tag{3}$$

The MN posteriors corresponding to the input feature sequence $\underline{\mathbf{x}}$: $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_t, \ldots, \hat{\mathbf{x}}_T \in B$ is estimated by Eqn.1, 2 and is minimized as a function of the network parameter $\theta = (f, g)$ using the CTC loss. This minimization is carried out over various sampling of (S, B) from a given training task T, i.e., S inherits all the class labels of the 'train' task T via a label set L (all phone classes/letters of the training set) sampled from T, and B defines the set of feature sequence-phone (or feature sequence-letter) ground truth pairs $(\underline{\mathbf{x}}, \underline{\mathbf{z}})$. The effectiveness of matching networks lies in this learning of $\theta = (f, g)$ as the embedding functions encapsulating the prior knowledge available in the training task T through the train support set S. This generalizability of (f, g), learnt from S on to S' is what gives the FSL advantage for matching networks in cross-domain operation.

3.2. MN-CTC Decoding

A test set T' is sampled to yield a label set L' from which the test query utterance $\underline{\mathbf{x}} \in B'$ (test query set) and test support set S' are sampled as in Figure 1. The test support set utterances are reconstituted by BiLSTM-CTC decoding from its underlying posteriors to yield the test support set S' for a N-way, K-shot decoding problem. Note that $\underline{\mathbf{x}} \notin S'$. The decoding proceeds with MN-CTC by first feeding input continuous speech $\underline{\mathbf{x}} : \hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_t, \ldots, \hat{\mathbf{x}}_T$ to derive the posterior sequence $\underline{\mathbf{y}} : \hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_t, \ldots, \hat{\mathbf{y}}_T$ through Eqns. 1 and 2 which is further acted upon by CTC prefix search decoding to yield the phoneme/letter sequence $\underline{\mathbf{z}} : \hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \ldots, \hat{\mathbf{z}}_m, \ldots, \hat{\mathbf{z}}_M$, $M \leq T$ on which the PER/LER is calculated with respect to the ground truth (phoneme/letter sequence) of $\underline{\mathbf{x}}$ from the test query set B'.

4. Cross-lingual MN-CTC scenarios

In this section, we establish and demonstrate the generalizability of the embedding functions f and g to cross-domain inference data as originally conceived in [4] to a cross-lingual scenario which constitutes our main contribution in this paper.

In this work, we validate the MN-CTC formulation for cross-lingual FSL through high-resource training of MN-CTC architecture (represented by the red annotated path in Figure 2) followed by low-resource adaptation of the source model using target language and inference on the target langauge test data (shown using the blue colour annotation in Figure 2). In cross-lingual MN-CTC framework, sufficient data as available in the source corpus is used for learning efficient embedding functions (f and g) which allows the support-set to have enhanced intra-class compaction and inter-class separability followed by inference of low-resource target languages, needing only very small data (few shots) from the inference-support-set during inference by few-shot-learning. The possible lack of alignment of cross-domain data - class distributions between source and inference languages is addressed by using a small amount of adaptation data of target language while re-training of the embedding functions (f and g) prior to inference on the target language.

In the following subsections, we present the MN-CTC architecture details for the cross-lingual operation and a transfer learning baseline against which we compare the MN-CTC results.

4.1. MN architecture details

In our MN-CTC formulation, embedding functions f and g are realized using the same deep neural network architecture ($f = g$). f and g are realized as 4 bidirectional LSTM (Bi-LSTM) layers of 512 cells, which maps each feature vector in an utterance (a sequence of 39-dimension MFCCs) to an embedding dimension of 1024. The embedding functions are learnt episodically from a train support set S generated from source language's utterances by repeated sampling in a P-way Q-shot manner. In inference, decoding of continuous speech test query utterances in B' is conditioned on support set S' formed by sampling the target language's validation utterances in a N-way K-shot manner.

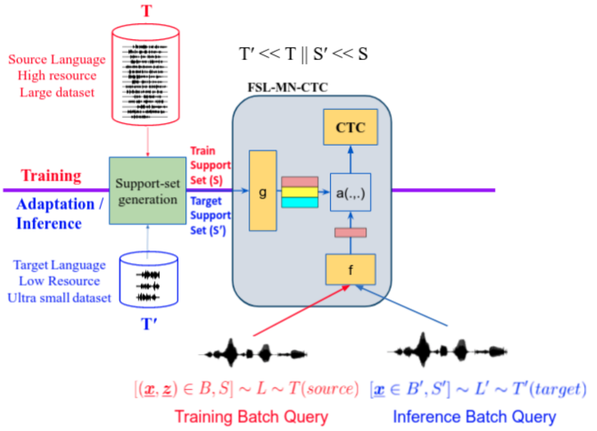

Figure 2: *MN-CTC cross-lingual framework*

4.2. Transfer learning baseline

MN-CTC results are compared against the traditional 'transfer learning' (TL) method - acoustic models trained on one domain/language is adapted to new languages which we refer in our paper as 'Baseline TL'. Here, in Baseline TL, an acoustic model is trained using Bi-LSTM-CTC network on high resource language (source language) and hidden layers are transferred from this pre-trained network on to the low resource training condition, the last fully connected layer of the source model is modified to meet the specifications (e.g. phone classes) of the the target language being trained and the entire network weights are fine-tuned with the target low resource training (validation utterances) data. The network is composed of 4 layer Bi-LSTM with each layer having 512 hidden units.

Here, for training the acoustic models we have used Adam optimizer with a learning rate 10^{-3} (or 0.001), momentum decay rate 0.9 and we choose the model that gives the best inference performance on the validation data-set.

4.3. Dataset details

In our experiments we use 6 Indian languages, namely Hindi, Gujarati, Marathi, Tamil, Malayalam and Kananda. Of these, the first 3 languages belong to the Indo-Aryan language family and the last three to the Dravidian language family. Hindi and Tamil speech corpus were obtained as part of the ASR challenge hosted by IIT Madras [19] and Google-Indic Gujarati, Marathi, Malayalam and Kannada are crowd-sourced high-quality multi-speaker speech data [20]. Details of these languages are given below and in Table 1 and 2.

Table 1: *Source Dataset details (SS: Support Set)*

Source Language	Duration	No. of utt	Train-set (SS & batch utt)
Indo-Aryan Hindi	50 hrs	33810	40 hrs
Dravidian Tamil	108 hrs	72074	104 hrs

Table 2: *Target Dataset details (SS: Support Set)*

Target	Language	Duration	No. of utt	Train-set (adaptation data - SS & batch)	Dev set (inference SS)	Test set (inference query)
Indo-Aryan	Gujarati	8	4272	15 min - 1 hr	1.05 hrs	47 min
	Marathi	3	1569	15 min - 1 hr	45 min	33 min
Dravidian	Kannada	8.5	4400	15 min - 1 hr	1.68 hrs	1.7 hrs
	Malayalam	5.5	4126	15 min - 1 hr	1 hr	31 min

1. **IITM Hindi** has 50 hours of data spoken by 8257 speakers comprising of 33810 sentences divided into train, dev, test set of 27131 (40 hrs), 3330 (5hrs) and 3349 (5hrs) sentences respectively.

2. **IITM Tamil** comprises of 108 hours of read speech obtained from 26938 native speakers with 72094 utterances split into train-set and evaluation-set having 69144 (104 hrs) ad 2950 (4hrs) utterances respectively.

3. **Google Indic Gujarati** contains around 8 hours of read speech acquired from 36 native speakers consisting of 4272 sentences.

4. **Google Indic Marathi** has 3 hours of speech data spoken by 9 native speakers with 1569 sentences.

5. **Google Indic Kannada** includes about 8.5 hours of read speech spoken by 59 native speakers consisting of 4400 utterances.

6. **Google Indic Malayalam** is 5.5 hours of read speech of spoken by 42 native speakers with 4126 sentences.

Here, Hindi and Tamil being large duration corpus are respectively used for training the Indo-Aryan and Dravidian source models of the baseline Bi-LSTM-CTC and MN-CTC architecture. The Indo-Aryan experiments utilizes Hindi train-set to train the source acoustic model and perform cross-lingual inference using Gujarati and Marathi as the target languages. The Dravidian cross-lingual experiments work with Tamil as the source language and Malayalam and Kannada as the target languages. The target language corpus is split into train, development and test set as shown in Table 2. MN-CTC cross-lingual inference makes use of the dev set as the inference support set (S') and the test set as the inference query (B').

From the train set of the target languages, we draw out utterances of different duration - varying from 15 min to 1 hour and use these as the adaptation data for re-training the baseline TL and MN-CTC source acoustic model. We work using such minimal amount of adaptation data during inference to conform to the MNs original cross-domain FSL theory, i,e. use large data for source model training followed by low-resource few-shot inference.

In our MN-CTC experiments, we use the notation P-way in training (source language) and N-way in inference (on target languages) and in cross-domain (across data-sets and classes) formulation of MN, N is different from P (used in training) and the test support set S' has classes different from the phone/grapheme classes in the train support set S. This detail is captured in Table 3. For e.g. Hindi source model uses $P = 55$ (comprising 54 phone classes and the blank '_' class) for the phone-recognition task, during adaptation using Gujarati target data, $N = 53$ (52 phonemes and the blank '_' class). The number of shots per class (few-shot examples) is fixed as 20 during training and inference (i.e., Q=K=20)

Table 3: *P-way / N-way details used in MN-CTC experiments*

Experiments	Source P-Way	Target N-Way	
Indo-Aryan family	Hindi	Gujarati	Marathi
PER	55	53	52
LER	68	75	64
Dravidian family	Tamil	Malayalam	Kannada
PER	41	58	51
LER	53	72	74

5. Results and Discussions

In this section, we present the results highlighting the advantage of MN formulation for cross-domain few-shot learning by applying MN-CTC to the cross-lingual scenarios, i.e., MN-CTC source model trained on a resource-rich 'source' language is used to infer low-resource 'target' language.

We carry out two distinct set of experiments, using the Indio-aryan and Dravidian languages, as each language family has its unique phonetic and acoustic characteristics [21], [22] and we show that the MN-CTC framework fares equally effectively under these two settings. Figure 3 illustrates the PER (plots in the top row) and LER (plots in the bottom row) results of our proposed cross-lingual MN-CTC network and Baseline TL for the two language families. The cross-lingual MN's advantage over Baseline TL is shown for varying amount of adaptation (target) utterances used in re-training the source model embedding function f and g. The adaptation utterance duration is varied from 15 minutes to 1 hour as shown in Table 2 (column 5) and is represented along the x-axis of the plots in Figure 3.

We now examine in detail the PER and LER performances from Figure 3 **A: Indo-Aryan** cross-lingual experiment (Hindi source model on Gujarati and Marathi target) and **B: Dravidian** cross-lingual experiment (Tamil source model on Malayalam and Kananda target).

We observe from Figure 3 (A: Indo-Aryan) the following:

- Cross-lingual MN-CTC model shows 13% (absolute) improvement in PER over the Baseline TL for Gujarati and 15% (absolute) for Marathi.
- For LER experiments, cross-lingual MN-CTC exhibits 7% (absolute) improvement in LER over the Baseline TL for Gujarati and 6% (absolute) for Marathi.

We observe from Figure 3 (B: Dravidian) the following:

- Cross-lingual MN-CTC PER outperforms Baseline TL by 3% (absolute) for Malayalam and 6% (absolute) for Kananda.
- Cross-lingual MN-CTC model shows 9% (absolute) improvement in LER over the Baseline TL for Malayalam and 4% (absolute) for Kannada.

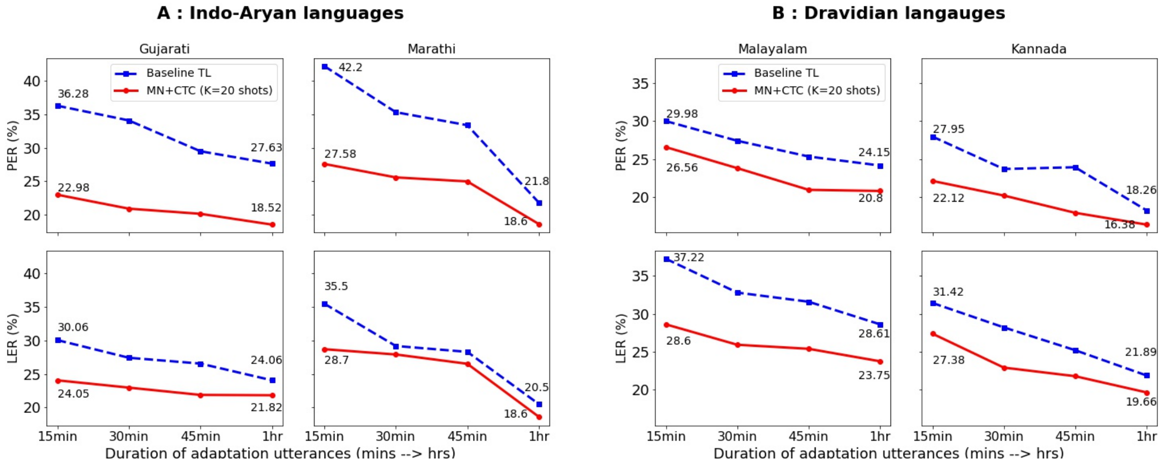

Figure 3: *Comparison between Baseline TL and crosslingual MN-CTC performance - PER (top row) and LER (bottom row) for Indo-Aryan (left) and Dravidian (right) languages*

We observe that our proposed Cross-lingual MN-CTC model consistently offers significant higher performance than the transfer learning baseline when target language adaptation data is as low as 15 min making it suitable for low-resource few-shot learning. As we increase the amount of adaptation data, the baseline TL shows a progressive performance gain though MN-CTC continuous to outperform the baseline TL at 1 hour of adaptation data also. This can be attributed to at least two reasons: Firstly, MN-CTC formulation essentially involves learning of embedding functions from a training support set which has phones (or classes, in general) which are not present in the test set and it is these embedding functions that generalize in a cross-domain manner to perform a KDE based classification of the unseen test samples. Secondly, this FSL framework allows dealing with unseen test phones (or classes, in general) not seen prior during training and represented only in the test support set requiring only a very few examples on which the test labeling needs to be done for both adaptation and inference.

These results emphasize the fact that FSL based matching networks can be used to realize low-resource cross-lingual speech recognition systems, where large source training data can be used to learn the appropriate f and g embedding functions which in turn impact the few-shot performance of a low-resource target language.

6. Conclusion

We have proposed the adaptation of a few-shot learning (FSL) framework ‘matching networks’ (MN) to the problem of E2E continuous speech recognition (CSR), in a first of its kind attempt. We have formulated a CTC-loss based end-to-end training of MN and an associated CTC-based decoding of input continuous speech. We present few-shot cross-lingual recognition by applying MN-CTC trained on a source language (resource-rich) to a target language (low-resource) thereby establishing the effectiveness of the MN-CTC FSL formulation for cross-domain few-shot learning.

7. Acknowledgements

We thank the MINRO (Machine Intelligence and Robotics) center at IIIT-B under which this work was carried out. This work was supported by Karnataka Innovation & Technology Society, Dept. of IT, BT and S&T, Govt. of Karnataka vide GO No. ITD 76 ADM 2017, Bengaluru; Dated 28.02.2018.

8. References

[1] Y. Woldemariam, Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic, Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 61-69, May 2020.

[2] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” ACM Computing Surveys, Vol. 53, No. 3, Article 63, June 2020.

[3] J. Lu, P. Gong, J. Ye, and C. Zhang,“Learning from very few samples: A survey,” Sept 2020.

[4] O. Vinyals, C. Blundell, T. Lillicrap, K.. Kavukcuoglu and D. Wierstra, “Matching networks for one shot learning,” In Advances in Neural Information Processing Systems (NIPS 16), pp. 3630-3638, 2016.

[5] E. Triantafillou, R. Zemel, and R. Urtasun, “Few-shot learning through an information retrieval lens,” In Advances in Neural Information Processing Systems, pp. 2255-2265, 2017.

[6] T. Pfister, J. Charles, and A. Zisserman. “Domain-adaptive discriminative one-shot learning of gestures,” In European Conference on Computer Vision. Springer, pp. 814-829, 2014.

[7] Tirthankar Banerjee, Narasimha Rao Thurlapati, V. Pavithra, S. Mahalakshmi, Dhanya Eledath and V. Ramasubramanian, “Few-shot learning for frame-wise phoneme recognition: Adaptation of matching networks,” in 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, August 2021.

[8] A. Graves, S. Fernndez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” pp. 369-376, 2006.

[9] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, ICML, 2014.

[10] S. Thomas, S. Ganapathy and H. Hermansky, “Crosslingual and multi-stream posterior features for low resource LVCSR systems,” In Proc. Interspeech, 2010.

[11] N. T. Vu and T. Schultz, “Multilingual multilayer perceptron for rapid language adaptation between and across language families,” In Proc. Interspeech, pp. 515-519, 2013.

[12] A. Ghoshal, P. Swietojanski, and S. Renals, “Multilingual training of deep neural networks,” In Proc. ICASSP, pp. 7319-7323, 2013.

[13] P. Bell, J. Driesen, and S. Renals, “Cross-lingual adaptation with multi-task adaptive networks,” In Proc. Interspeech, pp. 21-25, 2014.

[14] D. Chen and B. K. W. Mak, “Multitask learning of deep neural networks for low-resource speech recognition,” IEEE Trans. ASLP, vol. 23, no. 7, pp. 1172-1183, 2015.

[15] S. Tong, P. N. Garner, and H. Bourlard, “An investigation of deep neural networks for multilingual speech recognition training and adaptation” In Proc. Interspeech, vol. 2017-August, pp. 714-718, 2017

[16] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, “Sequence-Based Multi-Lingual Low Resource Speech Recognition, In Proc. ICASSP, vol. 2018-April, pp. 4909-4913, 2018.

[17] S. Tong, P. Garner, and H. Bourlard, “Multilingual training and cross-lingual adaptation on ctc-based acoustic model,” Speech Communication, vol. 104, 2017

[18] J. Hsu, Y. Chen, and H. Lee, “Meta learning for end-to-end low-resource speech recognition,” In Proc. ICASSP, Barcelona, Spain, pp. 7844-7848, 2020.

[19] “IITM Hindi Speech Corpus: a corpus of native Hindi Speech Corpus,” Speech signal processing lab, IIT Madras.

[20] F. He, S.-H. C. Chu, O. Kjartansson, C. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. Johny, M. Jansche, S. Sarin, and K. Pipatsrisawat, “Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems,”in Proceedings of The 12th Language Resources and Evaluation Conference (LREC), Marseille, France, pp. 6494-6503, European Language Resources Association (ELRA), May 2020

[21] A. Bakshi and S. K. Kopparapu, “Spoken Indian language identification: a review of features and databases, Sadhana, vol. 43, April 2018.

[22] Manjunath K E, Dinesh Babu Jayagopi, K. Sreenivasa Rao, and V Ramasubramanian, “Articulatory Feature based Methods for Performance Improvement of Multilingual Phone Recognition Systems using Indian Languages”, Sadhana (Springer), July 2020