Few-shot learning for End-to-End Automatic Speech Recognition

Dhanya Eledath
International Institute of Information Technology Bangalore (IIIT-B), India
dhanya.eledath@iiitb.ac.in

1. Key Challenges

Performance of ASR systems has improved substantially with the use of end-to-end trained deep learning models like Encoder-CTC [1], RNN-Transducer [2], Transformers [3] and Conformers [4], typically trained in supervised learning or self-supervised learning settings.

Supervised models require a significant amount of labelled data to estimate the network weights and to generalise to new test data, for e.g., the Whisper models [5] are trained on 680,000 hours of labeled audio data to attain SOTA WER ranging from 2% to 36% on various benchmark datasets. Since transcribing audio is prohibitively expensive and time-consuming, most languages have only a few hours of labelled speech available. As a result, most current-day ASR systems are limited to a small number of resource-rich languages such as English.

Lately, self-supervised learning has gained a lot of popularity wherein foundation models are 'pre-trained' on very large unsupervised speech data and used for downstream tasks by supervised fine-tuning and inference. The performance of downstream tasks is dependent on the amount of unlabelled data used for pre-training and model complexity [6]. More pre-training data (can range from 54,000 hours [7] to million hours [8]) yields better models and representations, but at the cost of high training time and computing power [6].

Learning paradigms of these systems are very different from the way humans learn: e.g., humans learn novel concepts from a handful of examples leveraging on previous experience. The newly emerging framework Few-Shot Learning (FSL) [9, 10] is a paradigm shift from prevalent large data requirements and seeks an alternative to learning new concepts from a few examples (as few as 1 to 5) per class during inference.

FSL methods belong to the class of meta-learning frameworks [11] wherein the prior knowledge acquired from different similar tasks is used to learn a new task quickly using very few shots per class. The strength of the FSL framework lies in a cross-domain training-inference scenario, where, for example, efficient transferable models (or embedding functions or representations) are learnt from a large training corpus in one domain; such learnt embedding functions are used as prior knowledge to perform few-shot inference in a possibly different domain and with classes not seen during training, potentially without any fine-tuning on target domain data.

2. Research Contributions

FSL methods have been applied to various tasks in computer vision, natural language processing [9, 10] and speech tasks such as rare-word recognition [12], sound event detection [13, 14, 15] and keyword spotting [16]. FSL methods can be broadly grouped into data-based, model-based and optimization-based approaches based on the learnt prior knowledge [9].

Model-agnostic meta-learning (MAML) [18], an optimization based approach (under 'algorithm as prior knowledge') is the only framework that has been adapted to E2E ASR under the ambit of FSL. MAML has been applied for tasks like multilingual speech recognition [19, 20, 21], code-switched speech recognition [22] and cross-accented speech recognition [23].

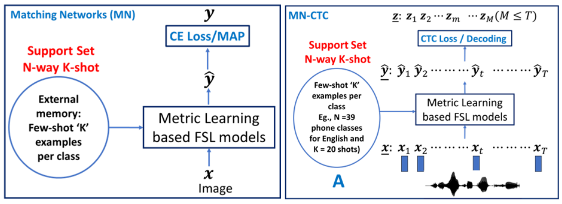
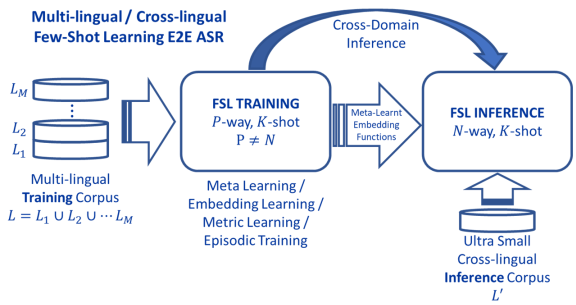


Figure 1: *Left: Original Matching Networks (MN) by Vinyals et al. [17]. Right: Proposed MN-CTC framework for E2E ASR*

In contrast to above MAML-based E2E ASR under 'algorithm-as-prior knowledge', my thesis focuses on exploring FSL techniques under 'model-as-prior-knowledge' for E2E ASR in a first-of-its-kind attempt, as outlined below in the form of a concise list of major contributions:

1. **Examine FSL based on 'model-as-prior knowledge'**
   I have focused on a classic and pioneering FSL framework - Matching Networks (MN) [17], simultaneously falling within the broad paradigms of Meta learning, Embedding learning and Metric learning within an Episodic Training or Sampling setting, to account for the matched condition between meta multi-tasks during training and inference, being defined as a N-way, K-shot FSL problem.
2. **In a first-of-its-kind attempt, adapt MN to E2E ASR**
   We first adapted the MN formulation (originally formulated for single image classification tasks as in Fig. 1, left panel) to frame-wise phoneme classification. This adaptation sets the basis for further applying the MN framework to continuous speech recognition. **[Publication 1]**

   For E2E ASR, I integrated MN into a Connectionist Temporal Classification (CTC) [24] loss based end-to-end training and CTC-based prefix-decoding of continuous speech in a network termed MN-CTC. **[Publication 2]**
3. **Apply cross-domain FSL definition to adapt MN-CTC to cross-lingual E2E-ASR**
   The primary characteristic of MN-CTC is the cross-domain applicability of the MN theory, where the test classes are different from the train classes. We applied MN-CTC to cross-lingual E2E ASR (Row 2 of Table in Fig. 2) for Indo-Aryan and Dravidian family of languages. By this, the proposed MN-CTC framework is highly effective for low-resource target languages yielding PERs/CERs far lower than conventional cross-lingual 'transfer learning'. **[Publication 2]**
4. **Major departure from data-hungry deep-learning trends**
   Matching Networks, set in a metric-learning FSL framework is a distance-based classifier. This intrinsically allows for few-shot (K-shots/class) training data in the form of non-parametrically represented support-set training vectors as external memory (marked "A" in Fig. 1, right panel). We show that MN-CTC - which we derive from this framework - needs



	Setting	M	L' vs L	Example
1	Mono-lingual	M=1	$L' = L$	English-to-English
2	Cross-lingual	M=1	$L' \neq L$	Hindi-to-Marathi
3	Multi-lingual	M>1	$L' \in L$	Indo-Aryan-to-Hindi
4	Multi-Cross	M>1	$L' \notin L$	Indo-Aryan-to-Tamil

Figure 2: *Top Panel: FSL pipeline involving cross-lingual inference on low-resource target language and Bottom Panel: scenarios arising from the above architecture*

as low as 15 min of data as inference support-set to perform cross-lingual inference on an unseen target language, and easily surpasses the performance of transfer learning frameworks under same few-shot conditions.

5. **Explore architectural variants of the MN-CTC network**
The labelled support set (annotated as 'A' in the right panel of Fig. 1) plays a crucial part in the MN framework during training and inference. For E2E ASR, we have proposed two architectural variants of MN-CTC for generating supervised support sets from continuous speech.
The first variant called 'Uncoupled MN-CTC' generates the support set 'outside' the MN-architecture and the second variant 'Coupled MN-CTC' generates the support set 'within' the MN-architecture through a multi-task formulation coupling the support-set generation loss and the main MN-CTC loss for jointly optimizing the support-sets and the embedding functions of MN. **[Publication 3]**

3. Methodology

3.1. Matching Networks (MN)

Matching Networks (MN) addresses the N-way K-shot FSL classification problem, where N (ways) is the number of classes and K (shots) is the number of examples per class. In the original MN framework (left panel in Fig. 1) by Vinyals et al. [17], the query (test sample) is an image sample ($\mathbf{x}$). MN embeds K-shot samples from N classes and the test sample $\mathbf{x}$ into a discriminative embedding space using embedding functions. Set in a distance-based classifier framework, MN converts the distances between $\mathbf{x}$ and the support set samples in the embedding space to a posterior estimate $\hat{\mathbf{y}}$, in a Neighborhood Component Analysis (NCA) [25] framework, used with cross-entropy (CE) loss for learning optimal embedding functions.

3.2. Adaptation of MN to E2E ASR

MN adaptation to E2E ASR using CTC loss (MN-CTC network) is depicted in the right panel of Fig. 1. Given an input continuous speech feature vector sequence $\underline{\mathbf{x}}$: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T$ and paired phone-label sequence ground truth $\underline{\mathbf{z}}$: $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m, \ldots, \mathbf{z}_M, M \leq T$, MN-CTC converts the distances between each $\mathbf{x}_t$ and the support-set samples to derive a posterior vector sequence $\underline{\mathbf{y}}$: $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_t, \ldots, \hat{\mathbf{y}}_T$ as required by the CTC loss function during training or by the CTC prefix-search decoding on test continuous speech. The test query utterance $\underline{\mathbf{x}}$ and support set samples are mapped by the learnt embedding functions to a highly discriminative space which allows classification of the query samples with very few labelled examples (K-shots).

4. Results and discussions

Here, we highlight few important results/observations inferred from the thesis directions and contributions discussed above.

4.1. Mono-lingual MN-CTC (Row 1 of Table in Figure 2)

Dataset: TIMIT [26] and Librispeech [27] corpus.
Observation: We realize low phone-error-rate (PER) / character-error-rate (CER) with the proposed MN-CTC yielding breakthroughs in very low data requirements ('K' shots, with K being as small as 10 to 20 frames per phoneme class).

4.2. Cross-lingual MN-CTC (Row 2 of Table in Figure 2)

Dataset: 1) Indo-Aryan case - Hindi [28] as the source language, Gujarati and Marathi [29] as targets. 2) Dravidian case - Tamil as source, Malayalam and Kananda [29] as targets.
Observation: Proposed Cross-lingual MN-CTC model offers a PER/CER advantage as high as 20-25% (absolute), over the transfer learning baseline for target language adaptation data as low as 15min, making it suitable for ultra low-resource E2E ASR. Table 1 shows the CER for varying target adaptation sizes.

Table 1: *Cross-Lingual MN-CTC CER results; Source Language - Hindi, Target languages - Marathi and Gujarati*

Target adaptation data size	**Marathi**		**Gujarati**	
	TL	MN	TL	MN
15 min	41.67	22.77	41.76	16.02
30min	35.48	20.61	29.34	14.8
45min	30.7	18.31	27.7	13.45
1hr	22.25	15.35	22.38	13.36
2.5hrs	13.73	10.54	18.44	10.93

4.3. Multi-lingual (Row 3 of Table in Figure 2)

Dataset: Indo-Aryan multi-lingual model trained on Hindi [28], Gujarati and Marathi [29]; Dravidian model on Tamil, Malayalam and Kananda [29]. Inference on target languages belonging to the respective family.
Observation: The Multi-lingual MN-CTC offers significant PER/CER performance advantage over a mono-lingual and cross-lingual MN-CTC, due to enhanced embedding functions learnt on phone classes with pooled multi-lingual data and consequent better generalizability to target languages.

5. Conclusion and Future Work

My primary focus is to bring in the strength of the cross-domain FSL property to E2E ASR and create a breakthrough in conventional high-resource settings, i.e., use 'ultra-low training data' FSL algorithms in place of current data-hungry deep learning systems. Future work will be along the following lines: 1) Explore 'Multi-cross' scenarios (Row 4 of the Table in Fig. 2) on Indo-Aryan and Dravidian language family to establish superior performance at ultra low-resource setting, 2) Set up FSL baselines like prototypical-network, relation-network, MAML based E2E ASR and non-FSL baselines like pre-trained foundation models.

6. 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.

7. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014.

[2] A. Graves, "Sequence transduction with recurrent neural networks," *ArXiv*, vol. abs/1211.3711, 2012.

[3] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[4] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *ArXiv*, vol. abs/2212.04356, 2022.

[6] A.-R. Mohamed, H. yi Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1179–1210, 2022.

[7] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020.

[8] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1519–1532, 2021.

[9] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, Jun 2020. [Online]. Available: https://doi.org/10.1145/3386252

[10] J. Lu, P. Gong, J. Ye, and C. Zhang, "Learning from very few samples: A survey," 2020.

[11] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 5149–5169, 2020.

[12] L. Florian and V. N. Thang, "Meta-learning for improving rare word recognition in end-to-end ASR," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5974–5978.

[13] Y. Wang, J. Salamon, N. J. Bryan, and J. Pablo Bello, "Few-shot sound event detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 81–85.

[14] K. Shimada, Y. Koyama, and A. Inoue, "Metric learning with background noise class for few-shot detection of rare sound events," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 616–620.

[15] S.-Y. Chou, K.-H. Cheng, J.-S. R. Jang, and Y.-H. Yang, "Learning to match transient sound events using attentional similarity for few-shot sound recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 26–30.

[16] H. Seth, P. Kumar, and M. Srivastava, *Prototypical Metric Transfer Learning for Continuous Speech Keyword Spotting with Limited Training Data*, 01 2020, pp. 273–280.

[17] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *NIPS*, 2016.

[18] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, 2017.

[19] J. Y. Hsu, Y. J. Chen, and H. Y. Lee, "Meta learning for end-to-end low-resource speech recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7844–7848, 2019.

[20] Y. Xiao, K. Gong, P. Zhou, G. Zheng, X. Liang, and L. Lin, "Adversarial meta sampling for multilingual low-resource speech recognition," in *AAAI Conference on Artificial Intelligence*, 2020.

[21] S. Singh, R. Wang, and F. Hou, "Improved meta learning for low resource speech recognition," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4798–4802, 2022.

[22] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning Fast Adaptation on Cross-Accented Speech Recognition," in *Proc. Interspeech 2020*, Oct 2020, pp. 1276–1280.

[23] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-transfer learning for code-switched speech recognition," in *Annual Meeting of the Association for Computational Linguistics*, 2020.

[24] 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*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 369376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[25] J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *NIPS*, 2004.

[26] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon Technical Report N, p. 27403, Feb. 1993.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 5206–5210. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178964

[28] IITM, "IITM Hindi Speech Corpus: a corpus of native Hindi Speech Corpus," *Speech signal processing lab, IIT Madras*, 2021. [Online]. Available: https://github.com/Speech-Lab-IITM/Hindi-ASR-Challenge

[29] 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: European Language Resources Association (ELRA), May 2020, pp. 6494–6503. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.800