Few-shot learning for E2E speech recognition: architectural variants for support set generation

Dhanya Eledath
International Institute of Information Technology Bangalore (IIITB)
Bangalore, India

Narasimha Rao Thurlapati
Samsung R&D Institute, Bangalore (SRI-B)
Bangalore, India

V. Pavithra
Samsung R&D Institute, Bangalore (SRI-B)
Bangalore, India

Tirthankar Banerjee
IIIT Bangalore
Bangalore, India

V. Ramasubramanian
IIIT Bangalore
Bangalore, India

***Abstract*—In this paper, we propose two architectural variants of our recent adaptation of a 'few shot-learning' (FSL) framework 'Matching Networks' (MN) to end-to-end (E2E) continuous speech recognition (CSR) in a formulation termed 'MN-CTC' which involves a CTC-loss based end-to-end episodic training of MN and an associated CTC-based decoding of continuous speech. An important component of the MN theory is the labelled support-set during training and inference. The architectural variants proposed and studied here for E2E CSR, namely, the 'Uncoupled MN-CTC' and the 'Coupled MN-CTC', address this problem of generating supervised support sets from continuous speech. While the 'Uncoupled MN-CTC' generates the support-sets 'outside' the MN-architecture, the 'Coupled MN-CTC' variant is a derivative framework which 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. On TIMIT and Librispeech datasets, we establish the 'few-shot' effectiveness of the proposed variants with PER and LER performances and also demonstrate the cross-domain applicability of the MN-CTC formulation with a Librispeech trained 'Coupled MN-CTC' variant inferencing on TIMIT low resource target-corpus with a 8% (absolute) LER advantage over a single-domain (TIMIT only) scenario.**

***Index Terms*—Few-shot Learning, Matching Networks, Continuous Speech Recognition, Coupled and Uncoupled architectures, Support Set Generation**

I. Introduction

The remarkable performance of deep learning architectures in the areas of speech, computer vision has been associated with large training data to learn complex models and generalize to unseen test conditions. This is very different from how humans learn. Humans are able to learn new concepts or recognize new objects from few examples and this has inspired new learning paradigms such as 'Few-shot Learning' (FSL) - which aims at learning novel categories with minimum supervision using few shots or examples per class [1], [2].

In our work, we focus on a metric-learning (embedding learning) based FSL paradigm - 'matching networks' (MN) [3] for the task of end-to-end (E2E) continuous speech recognition (CSR) using the Connectionist Temporal Classification (CTC) framework (termed MN-CTC) as first proposed by us in [4]. The central idea of MN is to use the learnt embedded knowledge to map the labeled class exemplars (support set samples) and unseen test (query) samples to a higher dimension space and predict the class label of the unlabeled query using a distance metric. In this paper, we re-introduce this formulation and propose architectural variants of the MN-CTC framework for efficient support set generation from continuous speech.

Matching networks framework classifies query samples conditioned on labeled support set examples. In image-recognition, class labels are available for individual images and dependency on a supervised support set does not pose a challenge [3]. In CSR, the training data comprises of feature-vector sequence and corresponding unsegmented ground-truth transcript; thus we need methods to generate frame level labels to form the supervised support set. CTC output being a distribution over all class labels and the 'blank' label [5], [6] requires the support set to have frames representing the blank class in addition to the phone class labels. To address the problem of support set generation (frame level labeled examples per phone class) from continuous speech for the fundamental operation of our proposed MN within CTC framework, we propose two different architectural variants:

1) 'Uncoupled MN-CTC' system that generates the supervised support set from continuous speech 'outside' the MN-CTC framework.
2) 'Coupled MN-CTC' system which combines the E2E MN-CTC training with a parallel support set generation pipeline 'within' the MN architecture in an end-to-end multi-task learning framework. In contrast to the uncoupled system, the coupled architecture generates the support set directly from continous speech in alignment with the embedding functions and offers superior performance at very few shots.

We further examine the advantage of the proposed 'Coupled MN-CTC' architecture for cross-domain adaptation. We use 100 hours Librispeech speech (high resource domain) corpus as source language to learn efficient embeddings in the phoneme/grapheme space and leverage the learnt embedding (prior knowledge) to infer on TIMIT (low-resource target language) with minimal adaptation data.

ISBN: 978-1-6654-6798-8

II. Relation to prior work

Few-shot learning methods have been widely used in areas of image classification [3], image retrieval [7], object tracking [8], gesture recognition [9] and language modeling [3]. Recently, FSL methods have been applied to various speech tasks like rare-word recognition [10], sound event detection [11]–[13] and continuous speech keyword spotting [14].

Recently, we adapted Matching Networks by Vinyals *et al.* [3] for frame-wise phoneme recognition [15], cross-lingual word recognition [16] and cross-lingual low-resource continuous speech recognition [4] using our MN-CTC framework. The MN-CTC framework in a first of its kind attempt, combines matching networks with CTC loss function for the task of E2E continuous phoneme recognition and constitutes the central part of the cross-lingual scenario [4].

In the current paper, we present two architectural variants of this MN-CTC framework - Uncoupled MN-CTC system represented in Fig. 1(a) and Coupled MN-CTC system represented in Fig. 1(b). The two networks differ in the approach used to generate supervised support sets (frame-wise labels of phoneme classes) from continuous speech required for the working of MN-CTC.

III. Matching Networks for E2E CSR

Matching Networks (MN) implement a N-way K-shot classification problem, where N (ways) is the number of classes and K (shots) is the number of examples per class. The main objective of MN is to embed the support-set and query samples to a discrimininative embedding space using appropriate neural networks (f and g) and use cosine-similarity measure between the support and query embeddings to classify the query to the correct support-set class.

We first re-introduce the basic theory of MN-CTC framework as proposed and elaborated by us in [4]:

1) **Supervised support set and query set:** Given a set of input-label pair (support set) and test sample (query), MN model classifies the query sample as one of the support set class labels. In MN-CTC network, the support-set includes all phonemes/graphemes of the language and input samples per class is generated from continuous speech by the methods we propose in Sec. IV. Query-set consists of continuous speech utterances and MN model embeds and classifies (in the embedded space) every frame in the query utterance to the corresponding support-set class.
2) **Metric Learning:** Input samples from the support set and the query set are mapped into an embedded space by the embedding functions f and g (realized using neural networks) respectively. In the embedded space, the similarity between the query sample and support set samples is estimated using the cosine similarity metric, constituting a metric learning.
3) **Episodic training:** To classify unseen classes during inference, MN network is trained in an episodic manner. In each episode, P classes with Q samples are randomly selected from a L label sampling of the training set (T) to form a P-way Q-shot support-set. Batch query set B consisting of continuous speech feature vector sequence $\underline{\mathbf{x}}$: $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_t, \ldots, \hat{\mathbf{x}}_T$ and the paired phone-label sequence ground truth $\underline{\mathbf{z}}$ is sampled from the training data T . Given the 'training' support set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$ (with $k = PQ$), the optimal embedding functions f and g are estimated by minimizing the CTC loss '$-\log P_\theta(\underline{\mathbf{z}}|\underline{\mathbf{x}})$' $\forall(\underline{\mathbf{x}}, \underline{\mathbf{z}}) \in B$ over various training episodes, i.e., the best network parameters $\theta = (f, g)$ are learnt episodically by minimizing the loss function in Eqn. (1) through back-propagation.

$$\theta = \arg\min_{\theta} E_{L\sim T}\left[E_{S\sim L,B\sim L}\left[\sum_{(\underline{\mathbf{x}},\underline{\mathbf{z}})\in B} -\log P_\theta(\underline{\mathbf{z}}|\underline{\mathbf{x}})\right]\right] \tag{1}$$

Episodic training which samples few examples per class matches the few-shot inference scenario and allows the MN model to generalize to new classes with few shots.
4) **Few-shot inference:** During inference, N classes with K samples are randomly selected from the validation set yielding a N-way K-shot support-set $S' = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$ (with $k = NK$). The test query utterance $\underline{\mathbf{x}}$ (from a test set, disjoint from S') and S' samples are mapped by the learnt embedding functions f and g to a discrimininative space (with enhanced intra-class compaction and inter-class separability of the classes) which allows to classify the query samples with very few labelled examples (K-shots) per class, i,e. the K value is as small as 10 to 20 frames per phoneme class.

IV. Architectural variants of MN-CTC

Labelled support-set examples play a crucial role in the MN model learning and in this work we address two different approaches to generate such labelled samples (frame-wise lables) from continuous speech utterances. Additionally, CTC formulation depends on blank labels for posterior sharpening and this necessitates the support set to include blank as a class in addition to phone-classes. In our work, we rely on BiLSTM-CTC decoded frame-level targets to yield frame-level labels and blank labels (a priori) in the support set. For this, initially we trained a BiLSTM-CTC network with paired continuous speech and unaligned phone/letter sequence (ground-truth) outside the MN framework and used the model to form a supervised support set. Here, the support-set preparation was taken care external to the MN formulation and we call this system as 'Uncoupled MN-CTC' as shown in Fig. 1(a). To reduce this effort in generating the support set explicitly as well as to optimally align the support set and the embedding functions, we designed a joint learning approach wherein the support set generation process was combined within the MN-CTC framework. This approach generates supervised support set samples along with MN-CTC training and decoding and we call this method as 'Coupled MN-CTC' as shown in Fig. 1(b).

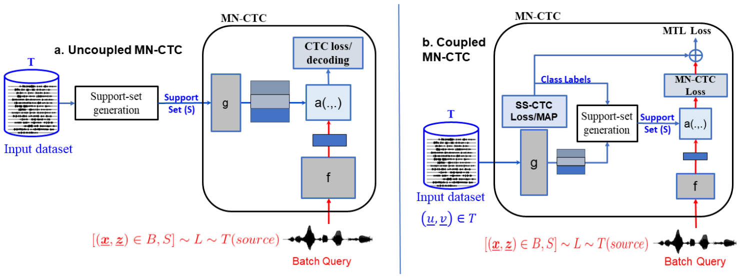

Fig. 1. a) Uncoupled MN-CTC and b) Coupled MN-CTC approaches to generate supervised support set

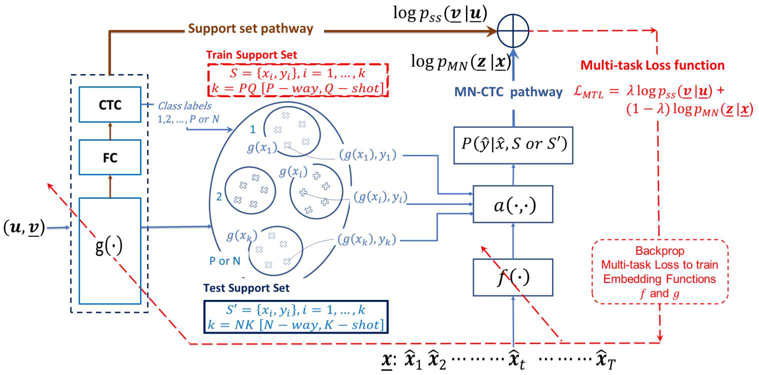
Support set pathway
$\log p_{ss}(\underline{v}|\underline{u})$
$\log p_{MN}(\underline{z}|\underline{x})$
Train Support Set
$S = \{x_i, y_i\}, i = 1, \dots, k$
$k = PQ\ [P - way, Q - shot]$
Multi-task Loss function
MN-CTC pathway
$\mathcal{L}_{MTL} = \lambda \log p_{ss}(\underline{v}|\underline{u}) + (1-\lambda)\log p_{MN}(\underline{z}|\underline{x})$
CTC
Class labels 1,2,…, P or N
FC
1
$g(x_1)$
$(g(x_1), y_1)$
$P(\hat{y}|\hat{x}, S\ or\ S')$
$g(x_i)$
2
$(g(x_i), y_i)$
$a(\cdot,\cdot)$
$(\underline{u},\underline{v})$
g(·)
$g(x_k)$
$(g(x_k), y_k)$
P or N
Backprop Multi-task Loss to train Embedding Functions f and g
$f(\cdot)$
Test Support Set
$S' = \{x_i, y_i\}, i = 1, \dots, k$
$k = NK\ [N - way, K - shot]$
$\underline{x}$: $\hat{x}_1\ \hat{x}_2 \cdots\cdots \hat{x}_t \cdots\cdots \hat{x}_T$

Fig. 2. Coupled MN-CTC approach to generate supervised support set

In an Uncoupled MN-CTC system, the support-set is generated outside the MN framework using a BiLSTM-CTC model. The $f()$ and $g()$ are trained only once using a CTC loss function from such an externally generated support set. In a Coupled MN-CTC system, the support generation is combined with MN training of $f()$ and $g()$ using a multi-task loss function trained end-to-end with only continuous speech as input.

A. Uncoupled MN-CTC

In this method, a BiLSTM-CTC is trained outside MN using the training set (T) utterances. The continuous speech utterances in the train (T) and test-set (T') are subjected to BiLSTM-CTC decoding to estimate the posteriors, on which a maximum a posterior (MAP) prediction yields frame-level labels which are then grouped phone-wise to form a non-parametric Q-shot cluster of each phone-class. In this way, the sequence of frames in an utterance are reconfigured into a P-way Q-shot representaion during training (N-way K-shot during inference) to form the support sets S (and S' respectively) as needed for the episodic support set sampling. The BiLSTM-CTC network output has a spiky posterior distribution and intrinsically produces blank labels for input frames near the boundaries. Frame-level CTC labels from BiLSTM-CTC decoded output include the blank labels, and provide the necessary support set samples for the blank class required for the operation of the MN-CTC framework [4].

B. Coupled MN-CTC

Training a BiLSTM-CTC outside MN requires an additional effort in collation of frames with phone labels and blank labels using MAP rule on the BiLSTM-CTC posteriors, which is cumbersome and time-consuming. So we resort to an efficient end-to-end multi-task learning framework integrated within the MN formulation called the Coupled MN-CTC. Here, the network architecture of embedding function 'g' includes BiLSTM embedding layer, a fully-connected (FC) layer and CTC loss-function as shown in Fig. 2. MN-CTC network is trained end-to-end to jointly optimize the support set loss and MN-CTC loss as in Eqn. (2).

$$\mathcal{L}_{\mathcal{MTL}} = \lambda \ \log \ P_{\text{SS}}(\underline{\mathbf{v}}|\underline{\mathbf{u}}) + (1-\lambda) \ \log \ P_{\text{MN}}(\underline{\mathbf{z}}|\underline{\mathbf{x}}) \qquad (2)$$

with tunable hyperparameter $\lambda : 0 \leq \lambda \leq 1$ that controls the task learning. Here, feature-label sequence $(\underline{\mathbf{x}}, \underline{\mathbf{z}})$ sampled from the batch B is fed to the MN-CTC pathway and $(\underline{\mathbf{u}}, \underline{\mathbf{v}})$ sampled from the training data T is fed to the support set pathway (brown solid line in Fig. 2) where $\underline{\mathbf{v}} : \hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_m, \dots, \hat{\mathbf{v}}_M$, $M \leq T$ is the paired phoneme/grapheme-label sequence ground truth of the input continuous speech feature vector sequence $\underline{\mathbf{u}}$: $\hat{\mathbf{u}}_1$, $\hat{\mathbf{u}}_2$, ..., $\hat{\mathbf{u}}_t$, ..., $\hat{\mathbf{u}}_T$. We then generate a supervised support of the embedded frames $g(\mathbf{x}_i)$ with labels y_i obtained from maximum a posterior (MAP) prediction of the CTC posteriors from the g network. g network, shared by both support set generation pathway and MN-CTC training/decoding pathway, optimizes the support set samples embedded by g to be aligned optimally with the batch utterances embedded by f. Coupled MN architecture works on paired data $(\underline{\mathbf{u}}, \underline{\mathbf{v}})$ of continuous speech $\underline{\mathbf{u}}$ and unaligned target phoneme/grapheme label sequence $\underline{\mathbf{v}}$.

V. Experiments and Results

A. MN-CTC configuration

In our MN-CTC formulation, embedding functions f and g are realized using the same deep neural network architecture ($f = g$ in a shared architecture). f and g are realized as four bidirectional LSTM (Bi-LSTM) layers of 512 cells, which maps each utterance (a sequence of 39-dimension MFCCs) to an embedding dimension of 1024. The embedding functions

TABLE I
PER/LER RESULTS OF MN-CTC (K = 20) AND BASELINE SYSTEMS

Architecture	TIMIT		Librispeech	
	PER	LER	PER	LER
Baseline systems				
Baseline-1 (all train utterances)	20.62	28.05	8.92	9.83
Baseline-2 (20 shots per class)	25.94	38.28	25.61	24.47
Proposed system (Q = K = 20 shots per class)				
Uncoupled MN-CTC	20.5	30.02	9.3	10.9
Coupled MN-CTC	19.7	28.3	9.8	10.6

are learnt episodically from a train support set S generated from train 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 validation utterances in a N-way K-shot manner.

B. Baseline systems

We compare our proposed MN-CTC systems with 2 different baseline systems:

1) Encoder-CTC trained using all the utterances in the train-set.
2) Encoder-CTC trained using the support-set samples (here, as small as 20 shots per class) exactly as seen by the MN-CTC in an episode. We carry out this experiment to understand the performance of standard deep-learning architectures in low-data conditions as established in [11].

The encoder network is realized as four bidirectional LSTM (Bi-LSTM) layers of 512 cells.

C. Dataset details

Our experiments are evaluated on two different English speech corpus; high resource Librispeech [17] and low-resource TIMIT speech dataset [18]. Librispeech consists of 1000 hours of read speech prepared from audio books as part of LibriVox project. In our work, we use 100 hrs of train-clean-100 as train-set, 5.4 hrs of dev-clean and test-clean as validation set and test set respectively. MN-CTC experiments on Librispeech corpus uses $P = N = 41$ for phone-recognition task (comprising 40 phone classes and the blank ('_') class) and $P = N = 28$ for letter-recognition task (including 26 letters, space and the blank ('_') class).

TIMIT speech corpus [18] contains 4.5 hours of phonetically-balanced read English speech. We worked using the standard train-dev-test split of the TIMIT database consisting of 4620 training utterances from 462 speakers, 500 validation utterances by 50 speakers and 240 utterances by 24 test speakers. Here, MN-CTC training uses $P = N = 40$ for PER experiments (comprising 39 reduced phone classes and the blank class) and $P = N = 28$ for LER experiments.

D. Results

Monolingual experiment: Here, we present the phoneme-error-rate (PER) and letter-error-rate (LER) to compare the two architectural variants of MN-CTC framework (Uncoupled MN-CTC and Coupled MN-CTC systems) discussed in this paper. The results obtained on test query utterances with a disjoint S' (support set formed from validation utterances) and K=20 shots are reported in Table I. We observe that

1) The coupled system outperforms the uncoupled system for both LER and PER experiments of TIMIT corpus. The two architectural variants offer comparable PER and LER results for Librispeech experiments.
2) The performance of our proposed systems compare very favorably to standard deep learning systems trained using large data (Baseline-1 in Table I).
3) The proposed systems offer far superior performance than the standard deep learning at very low (20 samples per class) FSL data sizes (Baseline-2 in Table I).

To understand the FSL advantage inherent in MN-CTC we also obtained the K-profile for S' = 500 utterances (25 min) shown in Fig. 3, for K = 6 to 80.

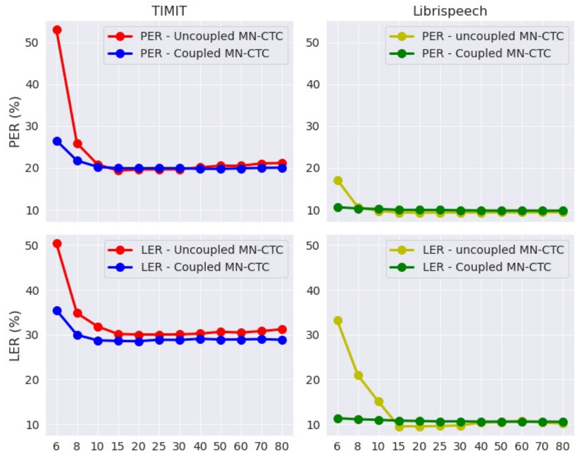

Fig. 3. PER/LER plot of TIMIT and Librispeech corpus for fixed S' = 500 utterances and varying K shot.

Small K (of the order of as low as 6-shots) yield $< 50\%$ PERs/LERs for TIMIT, and with increase in K towards 80, the MN-CTC offers progressively lower PER/LER. The K profiles of the coupled architecture are significantly better than the uncoupled architecture (for both PER and LER experiments) for very small shots (K = 6 to 10) and at higher shots both the variants offer comparable performance. The above results clearly establish that the multi-task learning in the coupled MN system jointly optimizes 'g' embedding and support set generation resulting in this performance advantage. Due to the superior performance of the Coupled MN-CTC architecture, we use this approach in our remaining experiments.

Cross-domain experiment: We examine continuous speech recognition under low resource conditions within a cross-domain adaptation of MN-CTC frame-work - i.e., training the network in a high-resource language, and applying it for decoding continuous speech of a 'low' resource target language, with few shots of input acoustic frames/class (phonemes/characters in the target language transcripts). To validate the MN formulation for cross-domain FSL, we use two different datasets of the same language English. The Coupled MN-CTC system is trained using Librispeech (high resource source corpus) and further used to decode the test utterances of TIMIT (low-resource target).

Sufficient data in Librispeech source dataset used for learning the embedding functions (f and g) effectively allows better generalizability of f and g to the target TIMIT inference and the support-set to have enhanced intra-class compaction and inter-class separability. By this, we need only very small data (few shots) from the inference-support-set during inference. The possible misalignment of cross-domain data (between source and inference domains) due to speaker variability, rendering style, channel differences is further addressed by re-training (adapting the embedding functions) using a small amount of adaptation data as illustrated in Fig. 4.

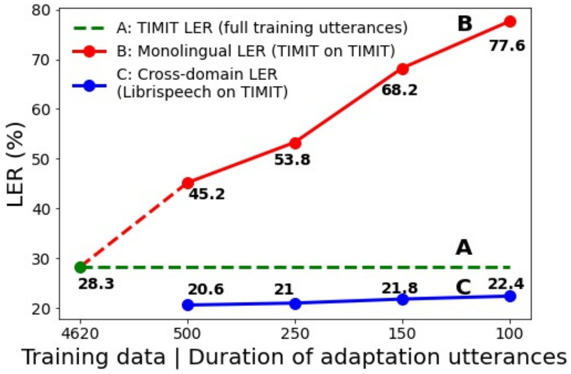

Fig. 4. Cross-domain training-inference performance (LER) of coupled MN-CTC network (K = 20 shots)

Fig. 4 represents the following sets of results:

1) Monolingual TIMIT LER for varying train utterances: **A**: MN-CTC model trained with 4620 utterances offers a baseline performance of 28.3% LER. **B**: Adaptation data of 500 to 100 utterances when used as monolingual training data for TIMIT yields a highly degraded baseline performance of 45.2% to 77.6% LER.
2) **C**: Cross-domain LER results of Librispeech (source) on TIMIT (target) for varying amounts of adaptation data - 500 (25 min) to 100 (5 min) utterances: We observe a 8% (absolute) improvement over the monolingual TIMIT for 25 minutes of adaptation data. This advantage remains almost invariant for adaptation data as small as 5 min (100 utterances).

VI. Conclusions

We have proposed and studied architectural variants of a FSL framework 'Matching Networks - CTC' (MN-CTC) for end-to-end (E2E) continuous speech recognition (CSR). The architectural variants are the 'Uncoupled MN-CTC' (the primary MN-CTC network) and the 'Coupled MN-CTC' which generates supervised support sets from continuous speech within the MN-architecture through a multi-task formulation coupling the support-set generation loss and the MN-CTC loss optimally. We have demonstrated the effectiveness of the proposed variants for 'few-shot' state of art PER and LER performances for CSR and also the cross-domain applicability of MN-CTC as a FSL paradigm using Librispeech source-domain and TIMIT target-domain in a low-resource setting.

VII. Acknowledgements

We thank the MINRO (Machine Intelligence and Robotics) center at IIITB 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.

References

[1] 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, 2020.

[2] Jiang Lu, Pinghua Gong, Jieping Ye, and Changshui Zhang, "Learning from very few samples: A survey," *arXiv:2009.02653*, 2020.

[3] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 3630-3638.

[4] Dhanya Eledath, V Pavithra, Narasimha Rao Thurlapati, Tirthankar Banerjee, and V Ramasubramanian, "Few-shot learning for cross-lingual end-to-end speech recognition," in *Workshop on Machine Learning in Speech and Language Processing 2021 (MLSLP 2021), Satellite Workshop of Interspeech*, Brno, Czech Republic, Sep 2021.

[5] Alex Graves, Santiago Fernndez, and Faustino Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning, ICML*, 2006, pp. 369–376.

[6] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. 31st International Conference on Machine Learning*, 2014, vol. 32, pp. 1764–1772.

[7] Eleni Triantafillou, Richard S. Zemel, and Raquel Urtasun, "Few-shot learning through an information retrieval lens," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[8] Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip H. S. Torr, and Andrea Vedaldi, "Learning feed-forward one-shot learners," in *Proc. 30th International Conference on Neural Information Processing Systems (NIPS'16)*, Red Hook, NY, USA, 2016, pp. 523–531

[9] Tomas Pfister, James Charles, and Andrew Zisserman, "Domain-adaptive discriminative one-shot learning of gestures," in *European Conference on Computer Vision. Springer*, Sep. 2014, pp. 814–829.

[10] Lux Florian and Vu Ngoc Thang, "Meta-learning for improving rare word recognition in end-to-end asr," in *2021 IEEE ICASSP*, 2021, pp. 5974–5978.

[11] Jordi Pons and Joan Serrà and Xavier Serra. "Training Neural Audio Classifiers with Few Data," in *2019 IEEE ICASSP*, (2019), pp. 16-20.

[12] Yu Wang, Justin Salamon, Nicholas J. Bryan, and Juan Pablo Bello, "Few-shot sound event detection," in *2020 IEEE ICASSP*, 2020, pp. 81–85.

[13] Kazuki Shimada, Yuichiro Koyama, and Akira Inoue, "Metric learning with background noise class for few-shot detection of rare sound events," in *2020 IEEE ICASSP*, 2020, pp. 616–620.

[14] Harshita Seth, Pulkit Kumar, and Muktabh Srivastava, *Prototypical Metric Transfer Learning for Continuous Speech Keyword Spotting with Limited Training Data*, 2020, pp. 273–280.

[15] 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, Aug 2021.

[16] Tirthankar Banerjee, Dhanya Eledath, and V Ramasubramanian, "Few shot learning for cross-lingual isolated word recognition," in *First International Conference on AI-ML-Systems (AIMLSystems 2021), Association for Computing Machinery*, New York, NY, USA, Article 15, pp. 1-7, DOI:https://doi.org/10.1145/3486001.3486235, 2021.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Queensland, Australia, April 2015, pp. 5206–5210.

[18] 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 Tech. Report N, Feb. 1993.