Few-shot learning for frame-wise phoneme recognition: Adaptation of matching networks

Tirthankar Banerjee
IIIT-Bangalore
Bangalore, India

Narasimha Rao Thurlapati
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

V. Pavithra
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

S. Mahalakshmi
Samsung R&D Institute, Bangalore (SRIB)
Bangalore, India

Dhanya Eledath
IIIT-Bangalore
Bangalore, India

V. Ramasubramanian
IIIT-Bangalore
Bangalore, India

***Abstract*—Recently, the topic of Few-Shot Learning (FSL) is emerging as a radical direction in machine learning, well established with a variety of paradigms and network realizations for image recognition. However, FSL is yet to emerge in speech recognition and allied topics. In this paper, we adapt an FSL paradigm 'matching networks' to the problem of speech recognition, in a first of its kind attempt, to different tasks such as multi-speaker small-to-medium vocabulary word recognition, monolingual and cross-lingual phoneme recognition tasks under mel-spectrogram and single-frame feature representations. The key to FSL is the ability to use extremely small number of training utterances, e.g. as low as 1 exemplar / class, as a N-way K-shot learning problem for large N and very small K (e.g. K=1 to 10). We show a remarkably high performance for each of the different speech recognition tasks considered here with matching networks, consistently requiring only very few 'shots' of exemplars/class, even while surpassing the performance of a direct application of KDE (kernel density estimation) without the matching network's embedding. This adaptation sets the basis for applying the matching network framework to continuous speech recognition and cross-lingual ASR with extremely low training requirements in the target test language.**

***Index Terms*—few-shot learning, matching networks, word and phoneme recognition, low resource, cross-lingual speech recognition**

I. Introduction

Traditionally, machine learning and deep learning paradigms have been associated with large training data considerations, specifically to leverage the underlying optimization aspects, such as in the estimation of large number of parameters (network weights) in deep-learning and to generalize to unseen test data adequately. In contrast, the recently emerging trend of 'Few-Shot Learning' (FSL) seeks an alternative to such large data paradigms and computational frameworks for learning, by pointing to the minimal 'training' data required by human cognitive mechanisms (i.e. a child learning a visual class such as a 'giraffe') without having to be exposed to millions of diverse examples of the class, essentially working towards the 'small data AI' scenarios. In addition to posing the baseline of human cognitive performance with limited training samples, FSL approaches are highly relevant in scenarios where supervised data is hard to get in large sizes, and any small size training paradigm (such as FSL) is bound to make an important difference, if it can offer performances comparable to large-data conditions.

This emerging topic of FSL has witnessed a wide variety of paradigms, theoretical frameworks and techniques and corresponding network realizations [1]. Few-shot learning framework, as the term implies, is defined to classify a test (unseen) sample (e.g. image) from 'few-shots', i.e. few examples per class, for instance, as few as 1 to 5; the specific case of having 1 example to learn and classify is referred to as '1-shot learning'. Being an emergent topic of research, FSL has been successfully formulated and applied to a certain classes of classification problems, e.g. image classification [2], image retrieval [3], object tracking [4], gesture recognition [5], language modeling [2]. In general, 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 the loss of minimizing empirical risk.

Specifically, we examine a 'model-based' (embedded learning) prior-knowledge paradigm termed 'matching networks' [2] for a class of speech recognition tasks. Importantly, our work is motivated by noting that the frameworks of FSL in general (and matching networks in particular), have not been addressed or applied to problems in speech recognition, where the large-data requirements for current deep-learning paradigms are intense, and any efforts to establish 'small-data' approches with acceptable performances can make a substantial difference, impacting the size of data required for realizing any given performance limit.

In our work, we examine a range of tasks such as word-recognition and phoneme-recognition under different input representations such as mel-filter spectrograms and single-frame (with context splicing) feature vector representations - leading to important implications for end-to-end continuous speech recognition, which we present in a companion submission [6] - where we propose an adaptation of the 'matching networks' framework within a CTC formulation for end-to-end training and continuous speech decoding; e.g. training the network in a possibly high-resource language, and applying it for decoding continuous speech of a target 'very low' resource language, with as little as 'few hundreds' of single-frame features of each phone class.

While noting that FSL has not been applied to speech recognition so far, we now contrast our work (based on matching networks for FSL) with a related work that comes closest to our objectives in speech recognition. Recently, in what can be viewed as a problem defined close to FSL, [7] adopt a zero-shot learning framework inspired by approaches in computer vision and propose the Universal Phonemic Model (UPM) to apply zero-shot learning to acoustic modeling. In relation to this work, our work attempts FSL in classifying unseen phonemes (not seen in the training support set), but calls for a very 'few-shot' examples in a test-support set defining the target task, which can also be a cross-lingual scenario.

II. Matching networks

The central theme of matching networks is to perform FSL during 'inference' by using a small set of K 'few-shot' samples (examples/class) to classify a 'test' sample within a posterior estimation method based on kernel-density estimation (KDE) and k-nearest neighbor (KNN) based classification [2]. This also belongs to metric-learning problem of learning the metric between the few-shot samples and test sample, which in matching networks takes the form of a cosine-similarity based attention mechanism. The attention form of metric further incorporates an embedding (of the few-shot samples and test sample), learnt during 'training', from prior knowledge in the form of 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. In the following, we first define the KDE-KNN classification generalization in matching networks, that forms the 'inference' part, followed by the actual matching networks 'training' formulation which learns the embedding functions constituting the network parameters.

A. Inference

We define the FSL 'inference' problem as a N-way, K-shot learning problem, wherein, a test sample $\hat{x}$ (e.g. an image) is to be classified as one of N class labels (visual objects), by using only the very few K-shot examples. This K-shot data is referred to as 'test support set' of k samples, which are input-label pairs $S' = \{(x_i, y_i)\}_{i=1}^{k}$, with $k = NK$. Here, y_i is one-hot encoded vector of dimension N. Rightfully, in conventional classification, this is the 'train' set of exemplars on which the classification depends on. Referring to this as the 'test' set is to distinguish from yet another 'train' set that matching networks use to learn 'prior-knowledge' in the form of embedding functions, further used in the FSL of N-way, K-shot classification of the test sample.

The inference part of FSL by matching networks maps the test sample $\hat{x}$ into a probability distribution over output labels $\hat{y}$ by performing the KDE/KNN generalization posterior estimate as in Eqn. (1), using the test support set S' of k samples.

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

Here, $\hat{y}$ is the set of N class labels the test sample can belong to. The posterior probability distribution over the set $\hat{y}$ is $P(\hat{y}|\hat{x}, S')$ from which the maximum a posteriori (MAP) prediction yields the class label to which $\hat{x}$ is classified as. Eqn. (1) gives the output for a new class as a linear combination of the labels in the test support set S', with the linear combination weights as the attention mechanism a, which is essentially a metric between $\hat{x}$ and each of the few-shot samples x_i in S'. In matching networks, $a(.,.)$ 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 $f = g$) to embed $\hat{x}$ and x_i respectively. The learning of the embedding functions f and g are dealt with under 'training' in the next sub-section.

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

This yields the inference 'with matching networks' as in panel **C** in Fig. 1). We also define the inference 'without embeddings' as the 'baseline system', i.e., when there is no embedding from the matching network training, as c operating directly on $\hat{x}$ and x_i. This is as in panel **B** in Fig. 1, which represents a non-FSL setting.

B. Matching networks training

As illustrated in panel **A** of Fig. 1, the matching network training involves learning the embedding functions f and g from a 'training' support set $S = \{(x_i, y_i)\}_{i=1}^{k}$ defined as a P-way, Q-shot set, i.e., with $k = PQ$. S comprises P classes not seen in the test-support S' and is hence referred to as the train-support set. Q is the number of examples per class in S. The matching network training finds the optimal network parameter $\theta = (f, g)$ by maximizing the objective function as in Eqn. (3).

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

Here, a batch $B = (x, y)$ provides the set of 'train' samples on which the log-likelihood of the posterior probability of the class label y of sample x, i.e., $P_{\theta}(y|x, S)$ as estimated by Eqn. (1), is maximized as a function of the network parameter $\theta = (f, g)$. This maximization is carried out over various sampling of S, B from a label set L drawn randomly (sampled) from a given training task T made of a super-set of class-labels, i.e., S inherits the class labels of the 'train' task T via a label set L sampled from T, and B defines the set of input-label pairs (x, y) likewise sampled (like S) from L, sharing the same label set as (but a set of samples x distinct from) S.

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, on to the inference in Eqn. (1) which in

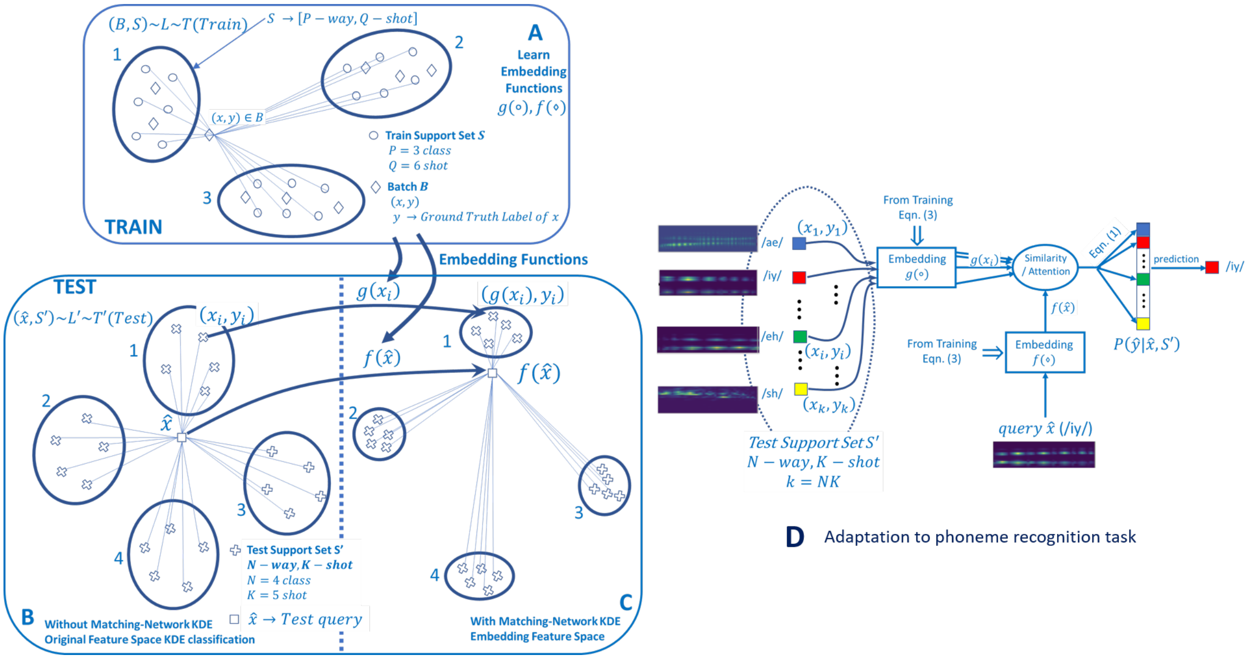


Fig. 1. *Matching Networks: Panel* **A** *- training using Eqn. (3); Panel* **B** *- inference* **without** *matching networks; Panel* **C** *- inference* **with** *matching networks; Panel* **D** *- Adaptation of the network to phoneme recognition task*

turn uses the attention mechanism derived through the cosine similarity metric c on the embeddings $f(\hat{x})$ and $g(x_i)$. This is illustrated in panel **C** of Fig. 1. Here it can be noted that this panel sets up the same inference as in panel **B** but in the embedding space $f(\hat{x})$ and $g(x_i)$ - resulting in highly enhanced class compaction (decreased intra-class variance) and increased inter-class separability. This generalizability of (f, g), learnt from S on to S' is what gives the FSL advantage for matching networks, a result we will see in the forthcoming sections for speech recognition tasks, particularly when S' is aligned with S in terms of having classes that share several types of similarity such as in terms of lower level abstractions (e.g. acoustic-phonetic features in the spectrographic representations) or in terms of classes that exhibit similarity, such as within words or broad phoneme categories or across languages.

III. Adaptation to speech recognition

We have adapted the matching networks formulation for a suite of speech recognition tasks, such as i) word level recognition from small to medium vocabularies from within TIMIT dataset, ii) phoneme recognition from within TIMIT and applied in a cross-lingual manner to another language Kannada (an Indian language) and iii) single-frame (frame-wise) phoneme recognition from within TIMIT and applied in a cross-lingual manner to Kannada.

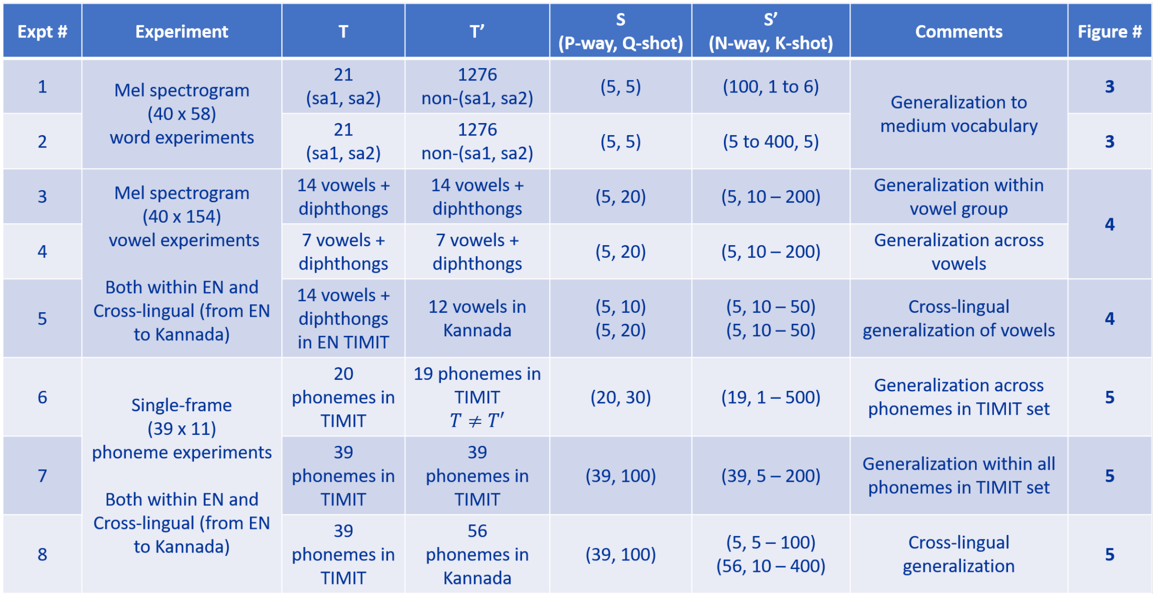
Our motive in adapting the FSL network to the above 3 sets of tasks is to a) establish that the FSL paradigm, possibly in a first attempt of its kind, applies to speech recognition tasks, both in a mono-lingual and cross-lingual setting, b) to show the high performance behavior of the matching networks framework to these tasks (for very few shots, in comparison to the large data requirements usually needed for realizing similar performances using current state of the art deep learning techniques) and c) the basis and potential for such a set of phoneme-level results to extend to continuous speech recognition [6], wherein the advantages of realizing low PER/WER systems by incorporating matching networks in the acoustic modeling and decoding frameworks leads to breakthroughs in the kind of 'very low' data requirements such systems will need with FSL incorporated.

The main adaptation in applying the matching networks to the speech recognition task involves using mel filter bank (FB) spectrogram representations a) (40 filter banks × 58 frames) for the word-level experiments and b) (40 filter banks × 154) for the phoneme experiments, with the variable length words/phones time-normalized using 'space-sampling' interpolation and warping [8] and c) (39 MFCCs × 11) single-frame feature representation for each frame centered within a ± 5-frame window.

The matching networks use a 4-layer convolutional neural network (CNN) for learning the embedding (f and g, with $f = g$); in each case the CNN takes the patch of 40×58 or 40 × 154 or 39 × 11 as input and yields a corresponding embedding in various higher dimensions in which the inference of Eqn. 1 is carried out as in Panel **C** in Fig. 1. As an example, Panel **D** in Fig. 1 shows the matching network for the phone-level experiment with the 40 × 154 mel filter bank spectrogram representation.

IV. Experiments and Results

We have conducted experiments on tasks as outlined in the table in Fig. 2. This table shows the type of different tasks and the associated data-set and definition of T, T', S, S' as relevant to the configuration of the matching networks. The corresponding results can be seen in the performance plots (%accuracy) in Figs. 3 to 5. We discuss the experiments and



Expt #	Experiment	T	T'	S (P-way, Q-shot)	S' (N-way, K-shot)	Comments	Figure #
1	Mel spectrogram (40 x 58) word experiments	21 (sa1, sa2)	1276 non-(sa1, sa2)	(5, 5)	(100, 1 to 6)	Generalization to medium vocabulary	3
2		21 (sa1, sa2)	1276 non-(sa1, sa2)	(5, 5)	(5 to 400, 5)		3
3	Mel spectrogram (40 x 154) vowel experiments Both within EN and Cross-lingual (from EN to Kannada)	14 vowels + diphthongs	14 vowels + diphthongs	(5, 20)	(5, 10 – 200)	Generalization within vowel group	4
4		7 vowels + diphthongs	7 vowels + diphthongs	(5, 20)	(5, 10 – 200)	Generalization across vowels	
5		14 vowels + diphthongs in EN TIMIT	12 vowels in Kannada	(5, 10) (5, 20)	(5, 10 – 50) (5, 10 – 50)	Cross-lingual generalization of vowels	4
6	Single-frame (39 x 11) phoneme experiments Both within EN and Cross-lingual (from EN to Kannada)	20 phonemes in TIMIT	19 phonemes in TIMIT $T \neq T'$	(20, 30)	(19, 1 – 500)	Generalization across phonemes in TIMIT set	5
7		39 phonemes in TIMIT	39 phonemes in TIMIT	(39, 100)	(39, 5 – 200)	Generalization within all phonemes in TIMIT set	5
8		39 phonemes in TIMIT	56 phonemes in Kannada	(39, 100)	(5, 5 – 100) (56, 10 – 400)	Cross-lingual generalization	5

Fig. 2. *Experimental scenarios: Matching networks for word and frame-wise phoneme recognition tasks*

results in the following with reference to this Table in Fig. 2. In all cases, the FSL advantage of matching networks can be noted, particularly in comparison to the very poor performance of the 'baseline' inferencing 'without embeddings' from the matching networks (as in Panel **B** in Fig. 1) and referred to as 'Baseline Cosine Similarity' in Figs. 3, 4 and 5.

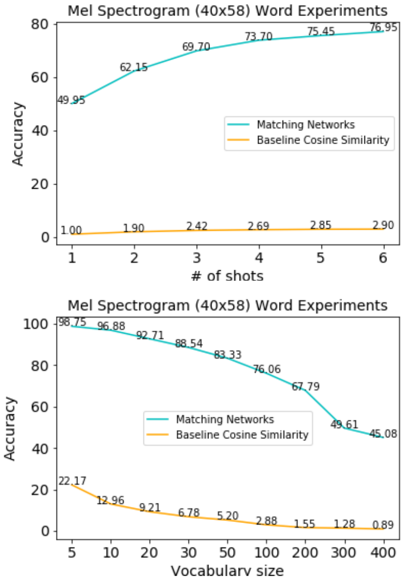


Fig. 3. *Expts # 1, 2: word-level experiments*

Mel FB spectrogram word-level experiments: Rows 1 and 2 show the configurations with T defined over 21 words from $sa1, sa2$ sentences in TIMIT, with S as $P = 5, Q = 5$ and made to generalize to a unseen set of 1276 words from non-$(sa1, sa2)$ sentences, for 2 cases - Fig. 3 (top) shows the accuracy for varying shots (1 to 6) for $N = 100$ sampled from the 1276 set and Fig. 3 (lower) shows the accuracy for varying vocabulary $N = 5$ to 400 sampled from 1276 words, but with fixed shots $K = 5$. The remarkable performance advantage of FSL matching networks (with embeddings) for very few shots (and increasing shots) over the non-FSL baseline can be noted.

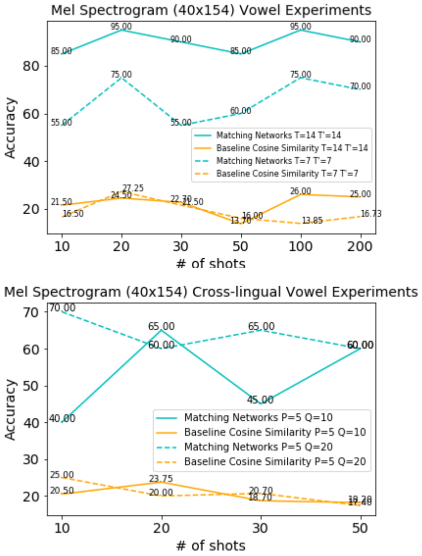


Fig. 4. *Expts # 3, 4 and 5: phoneme-level experiments*

Mel FB spectrogram phoneme-level experiments: Rows 3, 4 and 5 show the configurations for vowel-based experiments drawn from TIMIT with different T and T' definitions in each row. Row 3 shows the generalization 'within' the 14 vowel + diphthong set, and Row 4 shows the generalization across a 7-7 split of the 14 vowels representing an 'across' vowel generaliation. Fig. 4 (top) shows the accuracy variation for these 2 cases, with K-shots varying from 5 to 200 and the excellent FSL performance can be noted. In an important experiment for cross-lingual generalization of vowels, Row 5 shows T (and S) as TIMIT English vowels and T' (and S') as Kannada vowels. Fig. 4 (lower) shows the accuracy for this case, for varying K (10 to 50) for different $Q = 10$ and 20;

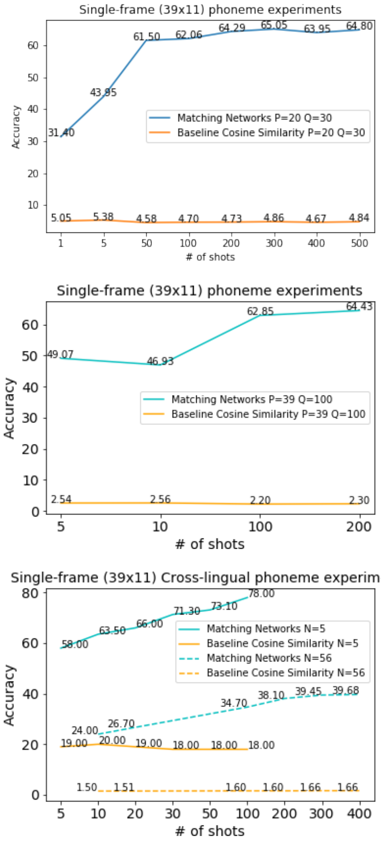


Fig. 5. *Expts # 6, 7, 8: single-frame phoneme-level experiments*

in each case the FSL advantage can be noted - and especially, the impact of increased Q on lifting the performance profile can also be noted. This points to the importance of the 'train' support set size P, Q in determining the generalizability of the learnt embeddings f, g.

Single frame phoneme-level experiments: Rows 6, 7 and 8 show the experiments of a different nature - that of using single frame representations for phoneme classification - both within TIMIT (English) and from TIMIT (En) to Kannada cross-lingual scenarios for all the phonemes of the datasets / languages. Row 6 shows a T, T' split of (20, 19) of the TIMIT reduced phone set of 39 phonemes - indicating 'across' phoneme generalization of the FSL learning. Fig. 5 (top) shows the accuracy for $S' = 19$ and K-shots 1 to 500, and Fig. 5 (middle) shows the accuracy for $S' = 39$ and K-shots 5 to 200. The FSL advantage can be clearly noted. Fig. 5 (bottom) shows the cross-lingual (English 39 phones to Kannada 56 phones) performance with K-shots varying. For S' as 5-way sampled from 56 Kannada phones, the performance profile is significantly high, while for S' as all the 56 phones, the performance profile understandably is lower, but nevertheless giving a response proportional to K-shots. Note that $K = 400$ single frame vectors correspond to a very low data condition of merely 4 secs of target Kannada sentences.

V. Discussion

Based on the above set of experiments and very consistent results of FSL advantage by the matching networks across word-level, phone-level and cross-lingual settings (particularly in comparison to the inference 'without matching networks'), these results reflect strongly in extending these scenarios to continuous speech recognition [6]. One of the pivotal information is the posterior probability vector ($P(\hat{y}|\hat{x}, S')$) for a single frame test vector $\hat{x}$; all of the results above clearly substantiate the significant 'sharpening' of this posterior probability vector by the matching networks and this in turn has important implications in all posterior based frameworks (e.g. Connectionist Temporal Classification or CTC formulations) in deep-learning based acoustic-modeling and decoding on the posterior vector sequence, for an input sequence of test vectors $\hat{x}$ as in continuous speech [6]. This would lead to applying FSL based matching networks for cross-lingual speech recognition, where a sufficient T and S (from a conventional large resource language) can be used to learn f, g which in turn impacts the performance for a few-shot inference in a target low-resource language defining T' and S'.

VI. Conclusions

We have adapted a few-shot learning framework 'matching network' for a suite of speech recognition tasks, by extending the framework to work with various representations at word-level, phone-level and cross-lingual settings. We show a remarkably high performance of each of the different speech recognition tasks, for the matching network FSL paradigm, consistently requiring only very few 'shots' of exemplars/class, even while surpassing the performance of a direct application of KDE (kernel density estimation) without the embeddings from the matching network. This adaptation sets the basis for extending the matching network paradigm to continuous speech recognition and for cross-lingual ASR with extremely low training requirements in the target test language.

References

[1] Yaqing Wang, Quanming Yao, James T. Kwok, 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.

[2] 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, Barcelona, Spain, 2016.

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

[4] L. Bertinetto, J. F. Henriques, J. Valmadre, P. Torr, and A. Vedaldi. Learning feed-forward one-shot learners. In Advances in Neural Information Processing Systems. 523-531, 2016.

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

[6] Dhanya Eledath, Narasimha Rao Thurlapati, V. Pavithra, Tirthankar Banerjee and V. Ramasubramanian. Few-shot learning for end-to-end speech recognition: Adaptation of matching networks. Submitted to EUSIPCO-2021, Dublin, Ireland, 2021.

[7] Xinjian Li, Siddharth Dalmia, David R. Mortensen, Juncheng Li, Alan W Black and Florian Metze. Towards Zero-shot Learning for Automatic Phonemic Transcription. Association for the Advancement of Artificial Intelligence (AAAI), 2020.

[8] S. Roucos, R. Schwartz, J. Makhoul. Segment quantization for very-low-rate speech coding. In Proc. ICASSP, vol. 3, pp. 1565-1568, 1982, Paris, France.