Curriculum Optimization for Low-resource Speech Recognition

Anastasia Kuznetsova[†], Anurag Kumar[⋆†], Jennifer Drexler Fox[⋆], Francis Tyers[†]

[†]Indiana University Bloomington [⋆]Rev.com, USA

Background

- Resource constrained ASR until now remains a challenging task. Our *curriculum learning* based approach mitigates the lack of labelled training data;
- A ranking function is applied to input data and then the data is split into K equal portions (tasks). The set of tasks is called *curriculum*.
- We use the ranking function as the prior on the input data, as well as the learner's progress to optimize the sequence of ASR inputs with the help of *multi-armed bandit* (MAB) algorithms.

Milti-Armed Bandit

- MAB is the concept from reinforcement learning (RL) domain. Here we describe its main components;
- The *agent* takes actions $a_k \in \mathcal{A}_K$ inside the environment; it selects the best action following the *policy* π;
- Policy π is a value function over all possible actions $\mathcal{A}_K$;
- The environment generates the *reward* r reflecting the optimality of the current action;
- The agent adjusts the policy based on the reward to improve its actions in the future.

Complexity metrics

- We hypothesize that the signal-based features have more significance for ASR than textual features and score input audios with *compression ratio* (CR) metric, where $Size_{before}$ is the size of the audio before the compression and $Size_{after}$ after compression:

$$CR = 1 - \frac{Size_{before}}{Size_{after}} \tag{1}$$

- Audios containing less noise compress more and have higher CR, on the other hand noisy audios have lower CR;
- We compare CR to text-based *sentence length* (SL) and *sentence norm* (SN) derived from the sentence embeddings;

References

Graves et al., "Automated curriculum learning for neural networks," in Proc. ICML'17.

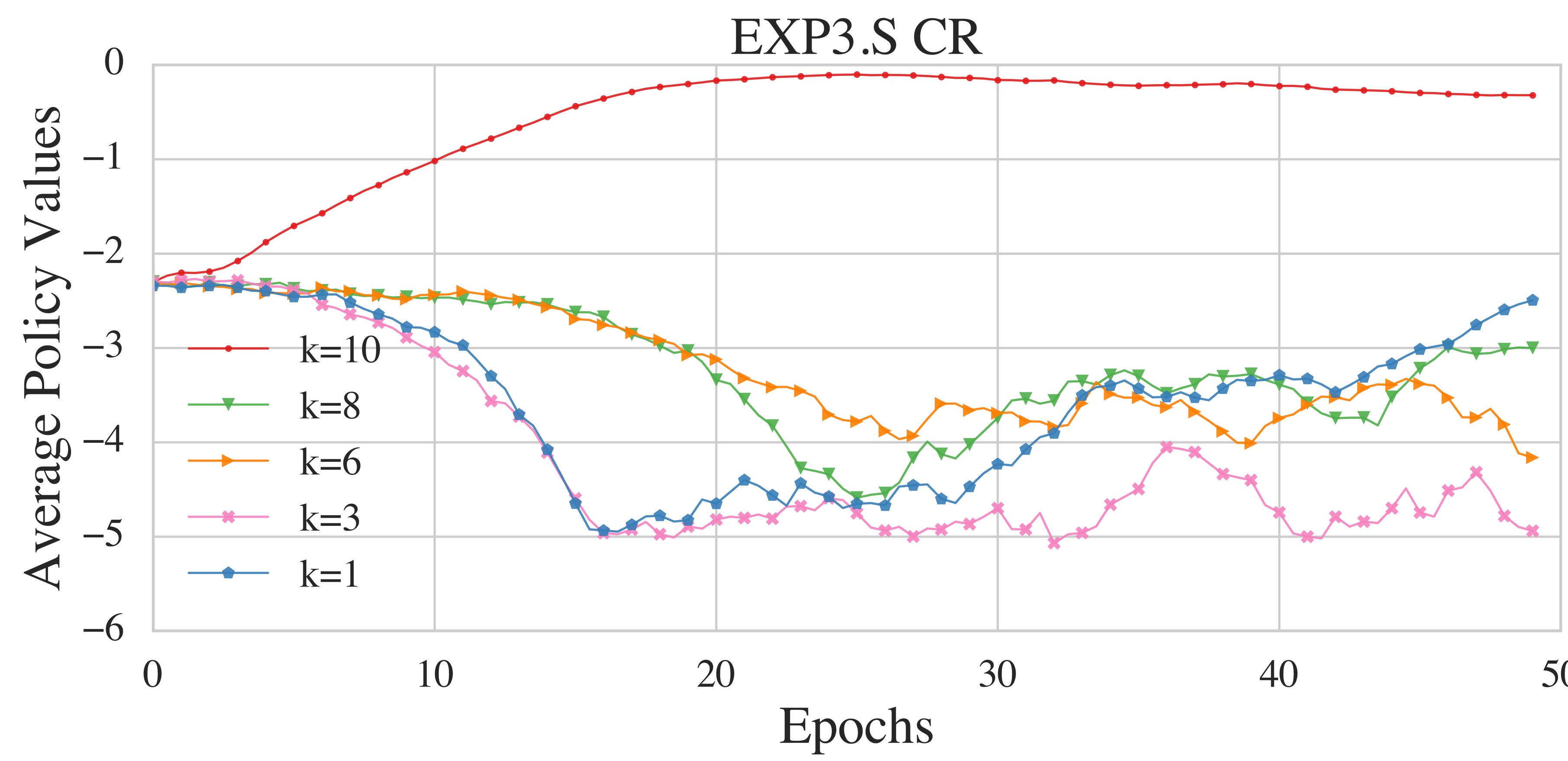

Figure: The figure shows log policy value change over time for selected k. This policy is generated by EXP3.S + CR on Kyrgyz (Ky) dataset and the policy is averaged per epoch.

	WER					CER				
Model	**Cv**	**Fy**	**Tt**	**Ky**	**Eu**	**Cv**	**Fy**	**Tt**	**Ky**	**Eu**
ESPNet + Trans	61.4	9.6	23.5	5.5	7.5	15.9	3.1	5.9	2.3	1.5
EXP3.S + CR	**41.8**	**7.8**	**22.3**	**4.1**	7.5	**9.8**	**2.6**	**5.6**	**1.9**	1.5
EXP3.S + SL	**41.1**	**9.3**	27.9	6.4	9.3	**9.7**	**3.3**	7.4	3.0	2.0
EXP3.S + SN	**42.7**	**7.8**	24.2	**4.8**	8.1	**10.2**	**2.5**	6.2	**2.2**	1.7
SW-UCB# + CR	**42.7**	**7.5**	**22.1**	**4.0**	8.0	**10.0**	**2.5**	**5.4**	**1.8**	1.6
SW-UCB# + SL	**42.4**	10.9	24.0	6.7	10.0	**9.9**	3.6	6.3	3.0	2.2
SW-UCB# + SN	**42.6**	**8.6**	24.8	**5.3**	8.6	**10.1**	**2.8**	6.4	**2.2**	1.8

Table: WER and CER for 5 selected languages. Baseline results are shown in the first row. Other rows show the combinations of the MAB algorithm and complexity metric, *CR* – Compression Ratio, *SL* – Sentence Length, *SN* – Sentence Norm for $K = 10$.

Experimental setup

- The ASR model is trained on 5 Common Voice 7.0 languages using ESPNet recipe; 80-dimensional log-mel filterbanks are extracted from the audio, with a window length of 25ms and a stride of 10ms; The data the is augmented with SpecAugment;
- The encoder has 9 Conformer blocks with 4 attention heads. The decoder has 6 transformer blocks. The network is trained with Adam optimizer with 25k steps warmup;
- The model is pretrained on Common Voice 7.0 English data set with the resulting WER of 15.2;
- The pretrained and fine-tuned on a target language models are used as a transfer learning baseline.

Method

- The input audios are ranked and split into K tasks with equal number of mini-batches;
- At each iteration t the MAB selects the best action k and updates the policy based on the reward r_t;
- The r_t is calculated from the loss-driven *self-prediction gain* (SPG)

$$\nu_{\mathrm{SPG}} = L(\mathcal{B}', \theta) - L(\mathcal{B}', \theta') \qquad \mathcal{B}' \sim D_k \tag{2}$$

where $\mathcal{B}'$ is the batch sampled from task D_k and θ are the parameters of the neural ASR model (Graves et al., 2017);
- We experiment with two MAB algorithms, probabilistic EXP3.S and deterministic SW-UCB# adapted for non-stationary problems.

Algorithm 1: Curriculum Learning

```
Initialize: D = f(X), π ← 0;
begin
    for t → T do
        Draw k based on current π;
        B_{t,k} ← sample(D_k);
        Train the model on B_{t,k};
        Observe progress gain ν_SPG;
        r_t ← g(ν_SPG);
        Update π on r_t;
    end
end
```

Results

- Our system achieved the highest relative decrease in WER of 33% relative for Chuvash (CV) and the minimum improvement of 5% for Tatar (Tt) language. For Fy, Tt and Ky, the best results are delivered with CR. For Cv the best WER was achieved by SL, however, the second best result is still achieved by CR.
- The policy over time (see Fig.) shows that harder $k = 10$ is preferred earlier in the training and policy value for easy task $k = 1$ increases towards the end of the training since the model gets more information from the harder examples in the beginning.

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2022.

{anakuzne, ftyers}@iu.edu {anurag.kumar, jennifer.drexler}@rev.com