

# CURRICULUM OPTIMIZATION FOR LOW-RESOURCE SPEECH RECOGNITION

Anastasia Kuznetsova,<sup>†‡\*</sup> Anurag Kumar,<sup>†\*</sup> Jennifer Drexler Fox,<sup>\*</sup> Francis M. Tyers<sup>‡</sup>

<sup>†</sup> Department of Computer Science, Indiana University Bloomington, USA

<sup>‡</sup> Department of Linguistics, Indiana University Bloomington, USA

<sup>\*</sup> Rev.com

## ABSTRACT

Modern end-to-end speech recognition models show astonishing results in transcribing audio signals into written text. However, conventional data feeding pipelines may be sub-optimal for low-resource speech recognition, which still remains a challenging task. We propose an automated curriculum learning approach to optimize the sequence of training examples based on both the progress of the model while training and prior knowledge about the difficulty of the training examples. We introduce a new difficulty measure called *compression ratio* that can be used as a scoring function for raw audio in various noise conditions. The proposed method improves speech recognition Word Error Rate performance by up to 33% relative over the baseline system.

**Index Terms**— speech recognition, curriculum learning, low-resource languages

## 1. INTRODUCTION

Automatic speech recognition (ASR) is the task of transcribing audio signals into text. Modern end-to-end ASR systems have achieved significant performance boosts over the past decade using various architectures and feature representations. However, ASR for low-resource languages still remains challenging. Conventional approaches to low-resource ASR such as transfer learning [1], multilingual ASR [2] or meta-learning [3] adhere to the notion of world languages sharing low-level components despite phonetic, structural and orthographic differences. Transfer learning uses the parameters of the model trained on a data-rich language to fine-tune the target language system, while multilingual ASR model is trained on several languages by sharing the hidden layer. Meta-learning addresses the problem of redundancy of language-specific features learned by multilingual models and further optimizes meta-loss to achieve language-independent acoustic representations. While these techniques improve WER, it is still much higher than in high-resource systems. More recent ASR approaches turned to self-supervised representations to reduce the need of labelled data while relying on large corpora of unlabelled audio [4]. We propose an

alternative approach achieving high recognition quality. Our model accounts for both the lack of training data and diverse noise/quality conditions based on the idea of *curriculum learning*.

*Curriculum learning* is a paradigm in supervised machine learning first introduced by [5]. It implies imitation of human learning strategies where the *curriculum* is a set of tasks organised in order of increasing complexity. Early curriculum learning strategies used a ranking function to determine the complexity of the data and organised the sequence of training examples according to this measure [5]. This is analogous to a teacher designed curriculum based on prior knowledge about the data, but it does not reflect the competence of the student over time. [6, 7, 8, 9] show that a curriculum yields better results with adaptive learning that accounts for the history of acquired competence during the previous time steps. [9] studied the dynamics of deep neural networks trained with dynamic instance hardness and showed that the model revisits harder samples more often due to higher variance in gradient values while easier examples tend to stay in the minima as soon as the minima are reached. However, [8] finds that in different empirical settings both learning harder and easy tasks first can benefit the model.

Curriculum learning has been successfully used in natural language processing tasks such as language modelling [5, 10] neural machine translation (NMT) [11, 12, 13], keyword spotting [14], and speech recognition [15, 16]. Most commonly two complexity strategies are employed: model competence-based [11, 12, 9] and data-driven. Complexity measures for data-driven learning include sentence/utterance length [16], language model score,  $n$ -gram size [5, 10], word frequency ranking, and sentence norm [13]. Speech processing systems can rely on speech to noise ratio (SNR) as a measure of difficulty by gradually blending more and more noise into clean speech signals [15, 14], but only if clean speech signals are available for creating noisy mixtures with different SNRs.

In this work, we propose a new complexity metric called *compression ratio* that helps to rank training examples based on audio quality even with the absence of clean data. Additionally, we demonstrate that both an external teacher curriculum and learner's progress are important for low-resource ASR. The teacher curriculum acts as a reliable prior while the student can construct its own curriculum in an automated

<sup>\*</sup>The work was partially completed while interning at Rev.com.

manner based on progress gains. We experiment with compression ratio and text-based difficulty measures to show that the signal-based prior leads to a more optimal solution.

The remainder of the paper is organized as follows: Section 2 describes the complexity measures and methods; Sections 3 and 4 discuss the experimental setting and the results; Section 5 summarizes our contributions.

## 2. METHOD

Curriculum learning for neural network optimization has been implemented in a variety of ways. Approaches in [11, 8, 9] propose a sampling strategy where a *scoring* function is the source of prior knowledge for the model obtained either from the properties of the data or from a pre-trained ‘teacher’ model and a *pacing* function that adjusts the sampling weights depending on the student’s progress. Our approach employs the *reinforcement learning* paradigm adopted by [10, 17]. Graves et al. [10] experiment with a k-armed bandit algorithm while [17] use more complex deep q-learning networks.

The reinforcement learning framework has three main concepts: *agent*, *environment* and *reward*. The *agent* is an entity residing in an environment; it selects the next best action following its *policy* (a value function over all possible actions). The *environment* generates the feedback reflecting the optimality of the current action. The environmental feedback is converted into a *reward* – a signal interpretable by the agent – and the agent adjusts its policy based on the reward in order to take better actions in the future. Formulating curriculum generation in terms of reinforcement learning allows us to incorporate both knowledge about the data and knowledge about the model’s learning progress. Sections 2.1 and 2.2 give a detailed description of curriculum formulation, task and complexity metric definitions.

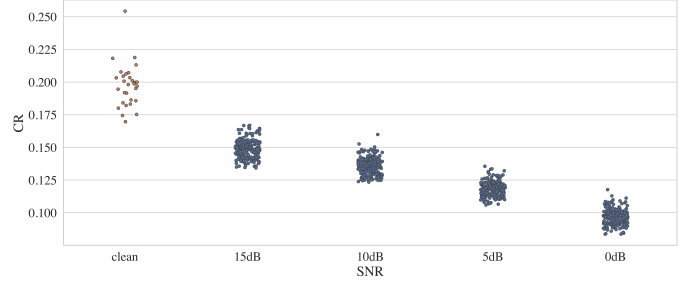
### 2.1. Complexity metrics

ASR systems receive a set of paired audio and text sequences as input. We can leverage both signal and text to build a scoring function. We hypothesize that, for speech recognition, acoustic properties have higher importance in guiding the neural network towards the optimum than textual features. Thus, we propose the *compression ratio* metric to score ASR input signals.

$$CR = 1 - \frac{\text{Size}_{\text{after}}}{\text{Size}_{\text{before}}} \quad (1)$$

Compression ratio is defined in (1) where  $\text{Size}_{\text{before}}$  is size of audio file before the compression and  $\text{Size}_{\text{after}}$  size after applying *gzip* compression algorithm respectively.

We assume that if there is less noise in the signal, the entropy of the signal will be lower and the audio will be more compressed with respect to its uncompressed version. On the other hand, noisier signals will be harder to compress and will yield lower compression scores. Figure 1 demonstrates this



**Fig. 1.** Compression ratio of clean audio compared to noisy mixtures at different SNRs (data from NOIZEUS[18]).

and shows that clean signals are compressed more as opposed to noisy mixtures of different SNR levels.

We also experiment with text based metrics such as the commonly used *sentence length* and *sentence norm*. Sentence norm is obtained from the vector representation of a sentence. Liu [13] notes that word vector norm is influenced by the properties of the word e.g. it increases with a decrease of the word frequency. Their experiments show that norm can be a good difficulty measure. We define sentence norm in (2) where  $y_j$  is a Word2Vec representation of BPE segment obtained from the *Sentencepiece* unigram model [19] and  $N$  is the length of the sentence.

$$SN = \left\| \frac{1}{N} \sum_{j=1}^N (y_j) \right\|_2 \quad (2)$$

Given the selected difficulty criteria, the *curriculum* is a set of training examples sorted in descending order by hardness and split into  $K$  subsets of data that we call *tasks*. Thus, we define a task set  $D = \{D_1, D_2, D_3, \dots, D_K\}$  where  $k$  is the index of the task and  $K$  is the empirically derived total number of tasks. Each  $D_k$  contains an equal number of training batches  $\mathcal{B}_{\text{bsize}}$  with the bsize (mini-batch size) of input audio.

### 2.2. Curriculum Learning Framework

---

#### Algorithm 1: Curriculum Learning

---

**Initialize:**  $D = f(X)$ ,  $\pi \leftarrow 0$ ;  
**begin**  
  **for**  $t \rightarrow T$  **do**  
    Draw  $k$  based on current  $\pi$ ;  
     $\mathcal{B}_{t,k} \leftarrow \text{sample}(D_k)$ ;  
    Train the model on  $\mathcal{B}_{t,k}$ ;  
    Observe progress gain  $\nu_{SPG}$ ;  
     $r_t \leftarrow g(\nu_{SPG})$ ;  
    Update  $\pi$  on  $r_t$ ;  
  **end**  
**end**

---

Our agent is defined as a  $K$ -armed bandit acting in the

reward space  $\mathcal{R}$  which aims to collect the maximum expected reward in the finite number of training steps. At each time step  $t$  based on the expected payoff the agent selects the best action which in our definition is equivalent to task index  $k$ . After selecting  $D_k$  the agent observes the reward  $r_t \in \mathcal{R}$  and updates the corresponding action value in policy  $\pi(k)$ .

The reward  $r_t$  is calculated based on the progress signals from the network termed *prediction gains*. We chose to experiment with *self-prediction gain*, a loss-driven progress signal proposed by [10]. SPG is an unbiased estimate w.r.t.  $D_k$ , it calculates the gain on a batch  $\mathcal{B}'$  sampled from the same task as the training batch  $\mathcal{B}$  (3).

$$\nu_{\text{SPG}} = L(\mathcal{B}', \theta) - L(\mathcal{B}', \theta') \quad \mathcal{B}' \sim D_k \quad (3)$$

Among the variety of  $K$ -armed bandit algorithms to optimize the curriculum we chose to experiment with stochastic *EXP3.S* algorithm proposed by [20] and tested on NLP tasks by [10], and the deterministic *SW-UCB#* algorithm by [21]. As both the environment and parameters of the ASR model are changing, the best task with the maximum payoff may change over time. Both of these algorithms are adapted for non-stationary problems; they ensure a balance between exploitation (choosing the best  $k$  at all times) and exploration (occasionally selecting a random  $k$ ) and minimize the regret in the long run. EXP3.S uses the  $\epsilon$ -greedy strategy to sample random  $k$  with  $\epsilon$  probability and select the best task with probability  $1 - \epsilon$ . The equation for policy update in EXP3.S is shown in (4), where  $K$  is the total number of tasks,  $t$  is the current time step and  $w$  is the weight for current  $t$  and  $k$ .

$$\pi_t^{\text{EXP3.S}} = (1 - \epsilon) \frac{e^{w_{t,k}}}{\sum_{j=1}^K e^{w_{t,j}}} + \frac{\epsilon}{K} \quad (4)$$

On the other hand, arm selection in SW-UCB# is based on the mean reward  $\bar{r}_k$  collected within a sliding window and the arm count within this time window.

$$k_t = \arg \max \{ \bar{r}_k(t-1, \alpha) + c_k(t-1, \alpha) \} \quad (5)$$

See (5), where  $\bar{r}_k$  is the estimate of mean reward for arm  $k$  within the sliding window,  $n_k$  is the number of times arm  $k$  was selected until time step  $t$  and  $\alpha$  controls the window size,  $c_k = \sqrt{\frac{(1+\alpha) \ln t}{n_k(t, \alpha)}}$ . In deep neural network training, the magnitude of the loss is higher in the initial stages of training and gradually decreases towards the convergence. To account for this, SW-UCB# has two environment modes, *abruptly-changing* and *slowly-varying*, with the mode determined by the variance of the recent loss history.  $\alpha$ , and thus the window size, is calculated differently in the two modes, so that smaller windows are used in the abruptly-changing mode.

Algorithm 1 shows a generalized version of Curriculum Learning. First, training data  $X$  is scored according to complexity function  $f(X)$ . Weight vector  $w$  and policy vector  $\pi$  (both of size  $K$ ) are initialized to 0. For each time step  $t$ , we select the best  $k$  based on  $\pi$ , then we sample batch  $\mathcal{B}_{t,k}$

	Eu	Fy	Ky	Tt	Cv
<b>Train</b>	45:08	18:25	24:17	19:27	1:41
<b>Dev</b>	8:03	4:14	2:07	2:47	0:45
<b>Test</b>	8:34	4:25	2:12	5:07	1:08

**Table 1.** Statistics for the datasets in hours and minutes.

from the corresponding task, train the model and observe the progress gain. Reward can be calculated directly from  $\nu_{\text{SPG}}$  or mapped by  $g(\cdot)$  to be in an interval  $r_t \in [-1, 1]$  to avoid the effect of high loss magnitude [10].

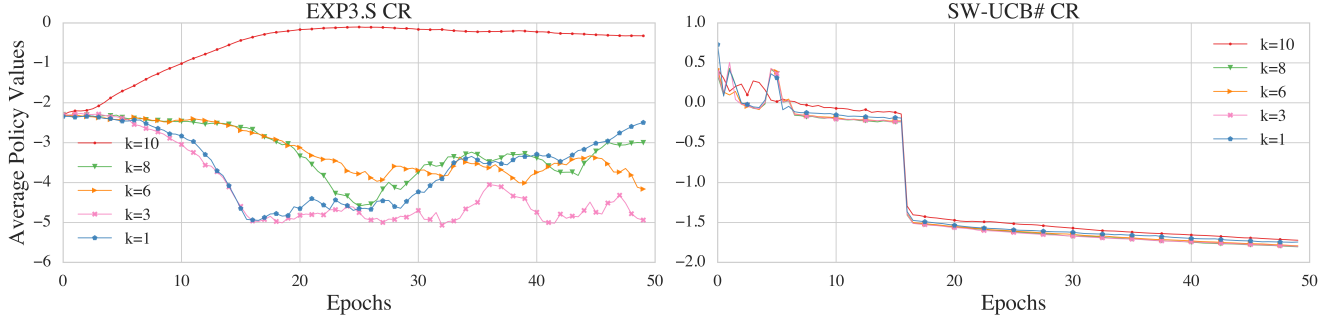
### 3. EXPERIMENTS

We evaluate the proposed approach on five low-resource languages from the multi-lingual crowd-sourced Common Voice 7.0 data set [22]: Basque (Eu), Frisian (Fy), Tatar (Tt), Kyrgyz (Ky) and Chuvash (Cv). The amount of training portion of the data ranges from less than 2 hours of audio (Cv) to 45 hours (Eu). Table 1 shows the statistics of the selected datasets.

#### 3.1. Experimental setup

All models were trained in ESPNet using the hybrid CTC /Attention model defined in the Common Voice recipe [23]. From the raw audio we extract 80-dimensional log-mel filter banks with a window length of 25ms and stride of 10ms. The data is then augmented using SpecAugment. We use Sentencepiece [19] to extract BPE segments from the transcriptions for RNN LM training. The acoustic model is comprised of an encoder that has 9 Conformer blocks with 4 attention heads and a decoder that has 6 identical transformer blocks. The network is trained with Adam optimizer with the warmup phase of 25k steps. In order to achieve better performance on low-resource data we apply transfer learning from an English ASR model pretrained on Common Voice 7.0 English corpus with a Word Error Rate (WER) of 15.2. Evaluation is done by averaging the 10 best checkpoints based on validation accuracy. We report Word Error Rate (WER) and Character Error Rate (CER).

In order to promote the most effective data usage we sample the audios with the highest policy values without replacement. This ensures data variability and better generalization of the model. Once the curriculum task with the highest value is exhausted, the training is continued using the rest of the data in order of decreasing policy values. The parameters used for SW-UCB# are  $\lambda = 12$ ,  $\gamma = 0.4$ ,  $\kappa = 0.8$ ,  $\nu = 0.1$  and for EXP3.S  $\epsilon = 0.05$ ,  $\eta = 0.001$ ,  $\beta = 0$  and a history size of  $10^3$  iterations.



**Fig. 2.** The figure shows log policy value change over time for selected  $k$ . This policy is generated by EXP3.S + CR and SW-UCB# combinations on Kyrgyz (Ky) dataset and the policy is averaged per epoch. The highest  $k = 10$  has larger value throughout the training while other  $k$  values are consistently lower.

Model	WER					CER				
	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	<b>41.8</b>	<b>7.8</b>	<b>22.3</b>	<b>4.1</b>	7.5	<b>9.8</b>	<b>2.6</b>	<b>5.6</b>	<b>1.9</b>	1.5
EXP3.S + SL	<u>41.1</u>	<b>9.3</b>	27.9	6.4	9.3	<u>9.7</u>	<b>3.3</b>	7.4	3.0	2.0
EXP3.S + SN	<b>42.7</b>	<b>7.8</b>	24.2	<b>4.8</b>	8.1	<b>10.2</b>	<u>2.5</u>	6.2	<b>2.2</b>	1.7
SW-UCB# + CR	<b>42.7</b>	<u>7.5</u>	<u>22.1</u>	<b>4.0</b>	8.0	<b>10.0</b>	<u>2.5</u>	<u>5.4</u>	<u>1.8</u>	1.6
SW-UCB# + SL	<b>42.4</b>	10.9	24.0	6.7	10.0	<b>9.9</b>	3.6	6.3	3.0	2.2
SW-UCB# + SN	<b>42.6</b>	<b>8.6</b>	24.8	<b>5.3</b>	8.6	<b>10.1</b>	<b>2.8</b>	6.4	<b>2.2</b>	1.8

**Table 2.** The table shows WER and CER for five selected languages. Results in bold indicate the improvement over the baseline, underlined values indicate best result overall. Baseline results with bare transfer learning are shown in the first row. The results below show the combinations of the algorithm and complexity metric, *CR* – Compression Ratio, *SL* – Sentence Length, *SN* – Sentence Norm for  $K = 10$ .

#### 4. RESULTS

We compare our approach to a transfer learning baseline trained with the parameters described in § 3.1. The results are presented in Table 2. Curriculum learning improves performance of the baseline ASR system with maximum  $\downarrow$ WER decrease of 33% for Cv and minimum improvement of 5%  $\downarrow$ WER for the Tt dataset. For Fy, Tt and Ky the best results are delivered by compression ratio metric. For Cv sentence length was more successful, however, the next lowest  $\downarrow$ WER is achieved with the combination of EXP3.S and compression ratio. The proposed approach gave better results in a very low data setting but did not improve on the larger Eu corpus.

Figure 2 shows the policy derived by the EXP3.S and SW-UCB# algorithms on Ky corpus. Policy values for EXP3.S start at the uniform distribution and then the hardest task  $k = 10$  gets the highest policy value in the early stages of training. It means that harder  $k$  is preferred earlier in the epoch and the model gets more information from the harder task. Task  $k = 1$  gradually increases the value towards the end of the training and the model selects easier batches in early iterations of final epochs. In early training during the abruptly-changing phase SW-UCB# switches between hard

and easy task and then it follows the same pattern as EXP3.S when the higher  $k$  is preferred continuing into the slowly-varying phase. The significant drop in SW-UCB# policy values happens when the training mode is changing from abruptly- to slowly-varying. This is explained by decreased reward and lower prediction gain magnitude, however this does not compromise the performance of the algorithm as  $\arg\max(\pi)$  of the policy is taken as opposed to sampling from the distribution in EXP3.S.

#### 5. CONCLUSION

Curriculum learning yields decreased WER in 4 low-resource languages. Our approach has shown that the derived curriculum based on the student’s learning progress improves performance and that the appropriate choice of complexity metric serves as a good prior for the model. Although text-based difficulty measures improve over the baseline model, CR outperforms them in several instances. We believe that the performance of curriculum learning is dependent on the complexity metric distribution and can be further optimized by selecting other values of  $K$ . We leave this study for the future work.

## 6. REFERENCES

- [1] J. Yi, Z. Tao, J. Wen, and Y. Bai, “Language-adversarial transfer learning for low-resource speech recognition,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 3, pp. 621–630, Mar. 2019.
- [2] J. Cho, M. Karthick Baskar, R. Li, M. Wiesner, S. Harish Reddy Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, “Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling,” *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 521–527, 2018.
- [3] Y. Xiao, K. Gong, P. Zhou, G. Zheng, X. Liang, and L. Lin, “Adversarial meta sampling for multilingual low-resource speech recognition,” in *AAAI*, 2021.
- [4] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *INTERSPEECH*, 2019.
- [5] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th ICML*, New York, NY, USA, 2009, ICML ’09, p. 41–48, Association for Computing Machinery.
- [6] Lu Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann, “Self-paced curriculum learning,” in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, AAAI’15, p. 2694–2700, AAAI Press.
- [7] T. Zhou and J. Bilmes, “Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity,” in *ICLR*, 2018.
- [8] G. Hach Cohen and D. Weinshall, “On the power of curriculum learning in training deep networks,” in *Proceedings of the 36th ICML*. 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544, PMLR.
- [9] T. Zhou, S. Wang, and J. Bilmes, “Curriculum learning by dynamic instance hardness,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 8602–8613, Curran Associates, Inc.
- [10] A. Graves, M. G. Bellemare, J. Menick, R. Munos, and K. Kavukcuoglu, “Automated curriculum learning for neural networks,” in *Proceedings of the 34th ICML - Volume 70*. 2017, ICML’17, p. 1311–1320, JMLR.org.
- [11] R. Wang, M. Utiyama, and E. Sumita, “Dynamic sentence sampling for efficient training of neural machine translation,” in *Proceedings of the 56th ACL (Volume 2: Short Papers)*, Melbourne, Australia, July 2018, pp. 298–304, Association for Computational Linguistics.
- [12] E. A. Platanios, G. Stretcu, O. and Neubig, B. Póczos, and T. Mitchell, “Competence-based curriculum learning for neural machine translation,” in *Proceedings of the NAACL 2019: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019, pp. 1162–1172, Association for Computational Linguistics.
- [13] X. Liu, H. Lai, D. F. Wong, and L. S. Chao, “Norm-based curriculum learning for neural machine translation,” in *Proceedings of the 58th ACL*, Online, July 2020, pp. 427–436, Association for Computational Linguistics.
- [14] Takuya H., Shreyas S., M. Souden, T. Tran, M. DeLafarah, and C. Dhir, “Dynamic curriculum learning via data parameters for noise robust keyword spotting,” in *Proceedings of ICASSP 2021*, 2021, pp. 6848–6852.
- [15] S. Braun, D. Neil, and S. Liu, “A curriculum learning method for improved noise robustness in automatic speech recognition,” *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 548–552, 2017.
- [16] K. Suyoun, M. L. Seltzer, J. Li, and R. Zhao, “Improved training for online end-to-end speech recognition systems,” *CoRR*, vol. abs/1711.02212, 2017.
- [17] M. Kumar, G. Foster, C. A. Cherry, and M. Krikun, “Reinforcement learning based curriculum optimization for neural machine translation,” in *NAACL*, 2019.
- [18] Y. Hu, “Subjective evaluation and comparison of speech enhancement algorithms,” *Speech Communication*, vol. 49, pp. 588–601, 2007.
- [19] T. Kudo and J. Richardson, “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” in *EMNLP*, 2018.
- [20] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, “The nonstochastic multiarmed bandit problem,” *SIAM J. Comput.*, vol. 32, pp. 48–77, 2002.
- [21] L. Wei and V. Srivatsva, “On abruptly-changing and slowly-varying multiarmed bandit problems,” in *2018 Annual American Control Conference (ACC)*. IEEE, 2018, pp. 6291–6296.
- [22] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” in *LREC*, 2020.
- [23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211.