Targeted Subset Selection for Limited-data ASR Accent Adaptation

Mayank Kothyari\textsuperscript{1,†}, Anmol Reddy Mekala\textsuperscript{1,†}, Rishabh Iyer\textsuperscript{2}, Ganesh Ramakrishnan\textsuperscript{1}, Preethi Jyothi\textsuperscript{1}

\textsuperscript{1} Indian Institute of Technology Bombay, Mumbai, India
\textsuperscript{2} The University of Texas at Dallas, Dallas, USA

{mayankk, anmolreddy}@cse.iitb.ac.in, rishabh.ry@utdallas.edu,
{ganesh, pjothy}@cse.iitb.ac.in

Abstract

We study the task of adapting an existing ASR model to a non-native accent while being constrained by a transcription budget on the duration of utterances selected from a large unlabeled corpus. We propose a subset selection approach using the recently proposed submodular mutual information functions, in which we identify a diverse set of utterances that match the target accent. This is specified through a few target utterances and achieved by modelling the relationship between the target and the selected subsets using these functions. The model adapts to the accent through fine-tuning with utterances selected and transcribed from the unlabeled corpus. We also use an accent classifier to learn accent-aware feature representations. Our method is also able to exploit samples from other accents to perform out-of-domain selections for low-resource accents which are not available in these corpora. We show that the targeted subset selection approach improves significantly upon random sampling - by around 5% to 10% (absolute) in most cases, and is around 10x more label-efficient. We also compare with an oracle method where we specifically pick from the target accent and our method is comparable to the oracle in its selections and WER performance\textsuperscript{1}.

Index Terms: speech recognition, asr personalization, data selection, accent adaptation, submodular mutual information

1. Introduction

State-of-the-art speech recognition systems have seen tremendous progress in the last few years, with end-to-end architectures becoming a default modeling choice. While end-to-end models yield impressive WERs and work well for certain user populations\textsuperscript{2}, they severely underperform when confronted with out-of-domain test utterances in target accents that are unseen or rarely seen during training\textsuperscript{3}\textsuperscript{4}.

A common solution\textsuperscript{3}\textsuperscript{5} to address such mismatched settings is to adapt a well-trained, speaker-independent ASR model with a small amount of target-specific data to adapt models to the target setting. While these works propose different fine-tuning schedules that would be most beneficial given the limited amount of target data, the question of which utterances should be chosen in order to be transcribed and further used for fine-tuning has received far less attention. Awasthi et al.\textsuperscript{6} present a method to select sentences within a fixed budget that are most likely to induce ASR errors and result in higher-quality personalized ASR models for target accents compared to random selection. However, they assume access to a small seed set of labeled utterances from the target speaker. We address the problem of selecting more realistic sets of utterances which we have access only to a limited number of unlabeled utterances from the target domain.

In this work, we propose a targeted subset selection approach that makes use of a suite of submodular mutual information (SMI) functions optimizing different objective functions (originally defined in\textsuperscript{8}). For a specific target accent we are given access to a few unlabeled speech utterances called the target set. We aim at identifying the most informative subset of speech utterances from a larger unlabeled pool that best matches the target set. Here, the best match is characterized by the underlying criteria of SMI functions, that makes use of pairwise similarities between speech representations. We find this to be an effective targeted subset selection technique for adapting ASR models in accents at multiple granularities within Indian accents and accents around the world. Our proposed subset selection method uses a limited transcription budget, i.e., around 10% of that of random. We even perform comparably to an oracle method that is privy to the target accent and selects only examples from this target accent.

2. Related Work

A number of works have studied subset selection for speech recognition.\textsuperscript{9}\textsuperscript{10} use submodular function-based subset selection on generated transcripts to find a minimal set of ASR training data and\textsuperscript{11} uses an entropy measure for the same.\textsuperscript{12} uses a joint Kullback-Leibler divergence-based subset selection on out-of-domain samples for ASR adaptation across acoustic characteristics such as speaker, noise and recording devices. Similarly,\textsuperscript{13} studies subset selection to obtain low-vocabulary speech corpora for ASR, while\textsuperscript{14} uses a submodular approach for targeted selection in machine translation. Many recent papers\textsuperscript{15}\textsuperscript{16} have studied uncertainty and gradient-based approaches for active learning to reduce the transcription time for ASR models, while\textsuperscript{17} uses a committee-based active learning method to select speech utterances.

A number of approaches have studied adaptation to atypical speech patterns like accented and dysarthric speech, such as\textsuperscript{5} and\textsuperscript{18} which fine-tune a subset of layers using labeled data from targeted accents.\textsuperscript{19} uses domain adversarial training to adapt across accents.\textsuperscript{20} works on a problem that corresponds exactly to the reverse of our setting by trying to determine the most error-prone sentences for the model to record utterances on. While this can be effective for user-driven personalization, our method is suited to settings in which we have fixed speech utterances, and the only actionable item for us is to transcribe a subset of them. All these approaches need data specifically labeled as coming from the target domain to use in fine-tuning.

Finally, a number of recent works have leveraged the submodular mutual information functions used in this work for targeted subset selection.\textsuperscript{21} use the SMI functions for query...
focused and privacy-preserving summarization, while [21] uses the SMI functions for improving the model performance on targeted slices. Recently, [22] proposed an active learning approach using the SMI functions for rare classes, redundancy, and OOD data.

3. Submodular Mutual Information (SMI) Functions

Submodular Functions: We let \( V \) denote the ground-set of \( n \) data points \( V = \{1, 2, \ldots, n\} \) and a set function \( f : 2^V \to \mathbb{R} \). The function \( f \) is submodular [23] if it satisfies the diminishing marginal returns, namely \( f(j | \mathcal{X}) = f(\mathcal{X} \cup j) - f(\mathcal{X}) \geq f(j | \mathcal{Y}) \) for all \( \mathcal{X} \subseteq \mathcal{Y} \subseteq V, j \notin \mathcal{Y} \). Submodularity ensures that a greedy algorithm achieves bounded approximation factor when maximized [24].

Submodular Mutual Information (SMI): Given a set of items \( S, T \subseteq V \), the submodular mutual information (SMI) [25, 8] is defined as \( I_f(S; T) = f(S) + f(T) - f(S \cup T) \). Intuitively, this measures the similarity between \( T \) and \( S \) and we refer to \( T \) as the targeted set. In the setting considered in this paper, the set \( T \) (target set) consists of a small set of unlabeled utterances from an accent, and \( V \) is a large unlabeled set of utterances from multiple accents. To find an optimal subset given a target set \( T \), we can define \( g_T(S) = I_f(S; T), S \subseteq V \) and maximize the same. Using a greedy algorithm, these submodular functions can be efficiently optimized within an approximation factor (1-1/e) of the global maximum.

3.1. Examples of SMI functions

We use the MI functions recently introduced in [8] and their extensions introduced in [23] [21]. For any two data points \( i \in V \) and \( j \in T \), let \( s_{ij} \) denote the similarity between them.

Graph Cut MI: The submodular mutual information (SMI) instantiation of graph-cut (GCMI) is defined as [21] [8]:

\[
I_f(S; T) = 2 \sum_{i \in S} \sum_{j \in T} s_{ij}
\]

Since maximizing GCMI maximizes the joint pairwise sum with the query set, it will lead to a summary similar to the query set \( Q \). GCMI models only query-relevance and does not select based on diversity [21].

Facility Location MI: The Facility Location Mutual Information (FLMI) function [21] takes the expression:

\[
I_f(S; T) = \sum_{i \in T} \max_{j \in S} s_{ij} + \sum_{j \in S} \max_{i \in T} s_{ij}
\]

FLMI is very intuitive for query relevance as well. It measures the similarity between representation of data points that are the most relevant to the query set, and vice versa.

Log Determinant MI: The SMI instantiation of logDMI can be defined as [21].

\[
I_f(S; T) = \log \det(S_S) - \log \det(S_S - S_{S,T}S_T^{-1}S_{S,T}^T)
\]

\( S_{S,T} \) denotes the cross-similarity matrix between the items in sets \( S \) and \( T \) and \( S_S \) denotes the similarity sub-matrix with rows and columns from the set \( S \). This function provides a trade-off between query-relevance and diversity.

4. Our Algorithm

Setting: We are provided a few unlabeled utterances from the accent (a target set \( T \)) which we would like the ASR model \( M \) to be adapted to.

Goal: The goal of the paper is to select the most informative subset \( S \) w.r.t. target \( T \) from a large corpus \( V \) of unlabeled data, called the ground set. We are given a budget constraint, which is a constraint on the total time of the selected utterances. This corresponds to the transcription budget, since the selected utterances need to be later transcribed by a human.

Approach: The approach is outlined in Alg. 1. The first step is to define the SMI functions. The SMI functions are defined by computing a euclidean similarity kernel \( s_{ij} \) on the features of the utterances (described in section 5). Using the SMI functions defined above, we select a subset that maximizes \( g_T(S) = I_f(S; T) \) for the given targeted set of utterances \( T \). Specifically, we optimize \( g_T(S) = I_f(S; T) \) subject to the constraint \( c(S) \leq B \), where \( c \) corresponds to the duration (in seconds) of the specific utterance, and \( B \) is the time budget. We use the greedy algorithm [24, 25] with a knapsack constraint for optimizing, where specifically given the current set \( S \), we select the item \( i = \arg \max_{s \in V \setminus c(S)} g_T(S) \), with the stopping criterion as \( c(S) \leq B \). Once, we obtain the set \( S \) as the solution of this optimization problem, we obtain the transcriptions from a human, and fine-tune the ASR model on the utterances in \( S \).

Algorithm 1: Adapting ASR model \( M \)

Require: target set \( T \), budget \( B \), SMI function type \( f \)
Data: large unlabeled ground set \( V \)
1 Define an SMI function \( g_T(S) = I_f(S; T) \) using Gaussian kernel similarity on the audio representation
2 \( \hat{S} \leftarrow \arg \max_{S \subseteq V, c(S) \leq B} g_T(S) \)
3 \( D \leftarrow \) Transcribe utterances in \( \hat{S} \)
4 Fine-tune ASR model \( M \) on \( D \)

5. Experimental Setup

Datasets: We experiment with adapting ASR models on two public datasets, viz., IndicTTS and CommonVoice containing English speech in both native and non-native accents. IndicTTS [27] consists of speech from 8 Indian speakers, each with a different accent depending on their native language: Gujarati (Gu), Kannada (Kn), Tamil (Ta), Malayalam (Ml), Hindi (Hi), Rajasthani (Ra), Assamese (As) and Manipuri (Ma). CommonVoice [28] contains samples from 12 accents across the globe including native accents like American, English, Canadian etc., non-native accents like Indian, African, Hong Kong and extremely low-resource accents like Bermuda and Malaysia.

For each accent, we created data splits by partitioning 70% of the data into a ground set \( (V) \) and a very small target set \( T \) (of sizes from 10 to 20) and we created test/dev sets from the remaining 30% using a 2/3 split, respectively. Figures 3 and 4 show the distribution of accented samples in our ground set for IndicTTS and CommonVoice, respectively, containing 34.5K and 33.8K samples overall. For IndicTTS, the average sizes of the ground set and the test set are 4.3K and 1.9K utterances, respectively. For CommonVoice, the sizes of splits vary widely across accents due to the ground set being biased (as seen in Fig. 4). Due to the small number of samples from
Bermuda and Malaysia, we treat them as low-resource accents and do not include any of their samples in the ground set.

**Feature representations.** We experiment with two feature representations for the speech utterances:

1. **MFCC features**: Each utterance is represented as a 39-dimensional feature vector of MFCC coefficients averaged over the duration of the utterance. Fig. 1 shows a t-SNE plot of MFCC features of IndicTTS test set utterances coloured by accent. We observed that accent-specific (effectively speaker-specific) clusters are well-separated from each other. However, this is because each accent uniquely maps to a different speaker. With larger speaker diversity within an accent, as in CommonVoice, MFCC features were no longer as discriminative across accents. This motivated the use of classifier-based features outlined next.

2. **Classifier-based features**: To generate the fixed-length vectors, we used the final layer of a classifier trained on audio samples from CommonVoice (disjoint from those used in submodular selection). We fine-tuned a wav2vec2-base [29] architecture (pretrained on LibriSpeech), freezing all except the last 2 encoder layers and then adding a mean-pooling and two dense layers to compose our classifier (see fig. 4). We used the last layer of this model as our accent embedding. This classifier achieved an accuracy of 83% accuracy when trained on 23K samples from 8 accents: US, england, scotland, ireland, indian, african, philippines and hongkong with around 2.9K samples per accent and tested on 200 samples per accent.

**ASR Model:** Following [7], our pre-trained model is based on the QuartzNet-15x5 [30] architecture. It is trained on LibriSpeech [31] for 400 epochs using the CTC-loss [32] and yields a WER of 3.90 on the test-clean split of LibriSpeech. The QuartzNet-15x5 architecture is fully convolutional with residual connections. This model is fine-tuned with our selected subsets of accented speech to minimize CTC loss using the NovoGrad optimizer [33] for 100 epochs with a batch size of 16, a linearly decaying learning rate of $10^{-5}$ and early stopping based on the dev set. In all our experiments, we report results averaged over three runs using three different seeds.

### 6. Experiments and Results

For each target accent we adapt to, we compare the following approaches: (i) random selection (baseline), (ii) SMI-based selection using three SMI functions, FLMI, GCMI and LogDMI and (iii) a skyline oracle system that uses the accent metadata and randomly samples from the targeted accent. Unless specified otherwise, we use target sets of size 10 and a transcription budget of 492 sec for IndicTTS, amounting to an average of 100 utterances. For CommonVoice we use a larger target size of 20, and a budget of 1080 sec (average of 200 utterances).
Table 1: WERs: subset selection on IndicTTS (B=492s).
Speaker: “accent, gender”.

| Speaker   | Pre-train | Random | FLMI | GCMI | LogDMI | Oracle |
|-----------|-----------|--------|------|------|--------|--------|
| Kn, M     | 18.73     | 15.79  | 14.07| 13.90| 14.77  | 14.20  |
| Mi, M     | 19.49     | 18.17  | 15.36| 15.73| 15.65  | 15.76  |
| Ra, M     | 21.88     | 16.87  | 15.71| 15.31| 15.33  | 15.58  |
| Hi, M     | 11.14     | 9.86   | 9.59 | 9.62 | 9.09   | 8.89   |
| Ta, M     | 12.48     | 11.88  | 11.79| 11.66| 12.26  | 11.92  |
| As, F     | 27.06     | 23.21  | 20.26| 20.40| 19.89  | 19.75  |
| Gu, F     | 13.73     | 11.21  | 9.98 | 10.40| 9.83   | 10.17  |
| Ma, F     | 53.08     | 46.98  | 41.78| 42.69| 42.39  | 42.15  |

Table 2: WERs: subset selection on CommonVoice (B=1080s).

| Accent      | Pre-train | Random | FLMI | GCMI | LogDMI | Oracle |
|-------------|-----------|--------|------|------|--------|--------|
| indian      | 43.73     | 40.86  | 38.43| 38.29| 38.99  | 38.22  |
| philippines | 30.45     | 29.83  | 29.28| 29.3 | 28.91  | 28.87  |
| hongkong    | 32.15     | 31.98  | 28.6 | 29.22| 30.28  | 28.65  |
| african     | 23.75     | 23.22  | 22.55| 22.59| 22.54  | 22.46  |
| scotland    | 51.72     | 50.07  | 46.31| 46.32| 46.42  | 45.42  |
| malaysian   | 35.59     | 34.92  | 34.64| 33.76| 34.25  | -      |
| bermuda     | 23.91     | 24.15  | 23.75| 23.48| 23.46  | -      |

Fig. 5 shows the absolute WER improvements using GCMI, FLMI and LogDMI for Assamese (IndicTTS), Indian and Hong Kong accents (CommonVoice) by varying durations of the selected subsets. We see that the SMI functions are able to achieve the performance of random selection with only around 10% of the data or lesser. Note that random selections perform better on Indian and Assamese than Hong Kong only because the first two target accents are well represented in the ground set (see sec. 3.1), and this explains its poorer selections from the target domain.

Table 4: Sample counts of accents picked by SMI functions when Wav2Vec2-based MFCCC was targeted from CommonVoice, B = 1080s.

| Function | FLMI | GCMI | LogDMI | Random |
|----------|------|------|--------|--------|
| WER       | 195/203| 187/194| 127/223| 102/224|
| Selections| 38.43| 38.29| 38.99| 39.36|

Adaptation results: Tables 1 and 2 show our main results on adaptation to target accents in IndicTTS and CommonVoice, respectively. We see consistent WER reductions using all three SMI functions compared to the random selection baseline. Our SMI-based systems perform comparatively to the skyline which picks utterances exclusively from the target accent.

Both FLMI and GCMI are able to almost exclusively select speech samples from the target accent, and LogDMI also picks majority samples in the selected subset from the target accent (see table 4). In Table 3, we see that better selections yield better WERs; random performs poorly and LogDMI worse than the other SMI functions due to not picking enough number of examples from the target accent. We see in Table 3 that using features that are poor in discriminating across accents lead to worse WERs. LogDMI compromises query coverage more for the sake of diverse selections (see sec. 3.1), and this explains its poorer selections from the target domain.

Table 4: Sample counts of accents picked by SMI functions when Indian was targeted (from CommonVoice), B = 1080s.

In Table 4, we also show accent adaptation to Malaysia and Bermuda English accents that were not available in the ground set. With the help of the classifier-based features, the SMI functions are able to use samples from other related accents for effective adaptation. (The oracle number is empty for these two accents due to their unavailability in the ground set.)

Fig. 5 shows the absolute WER improvements using GCMI, FLMI and LogDMI for Assamese (IndicTTS), Indian and Hong Kong accents (CommonVoice) by varying durations of the selected subsets. We see that the SMI functions are able to achieve the performance of random selection with only around 10% of the data or lesser. Note that random selections perform better on Indian and Assamese than Hong Kong only because the first two target accents are well represented in the ground set (see figs. 4, 5) compared to Hong Kong.

7. Conclusions

In this work, we propose a targeted subset selection approach using submodular mutual information functions that help identify speech utterances within a limited budget that best match an unlabeled target set containing accented speech. Our method consistently outperforms random selection methods and is comparable to an oracle system when the target set is drawn from a specific accent. Future extensions of this work will focus on extending this approach to deal with out-of-domain selections across low-resource accents by utilizing unlabeled data.
8. References

[1] K. Rao, H. Sak, and R. Prabhavalkar, “Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer,” in 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017, pp. 193–199.

[2] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina et al., “State-of-the-art speech recognition with sequence-to-sequence models,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4774–4778.

[3] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying bias in automatic speech recognition,” arXiv preprint arXiv:2103.15122, 2021.

[4] A. Koennecke, A. Nam, E. Lake, N. Dudell, M. Quatrey, Z. Menge-sha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, “Racial disparities in automated speech recognition,” Proceedings of the National Academy of Sciences, vol. 117, no. 14, pp. 7684–7689, 2020.

[5] J. Shor, D. Emanual, O. Lang, O. Tuval, M. Brenner, J. Cattiau, F. Vieira, M. McNally, T. Charbonneau, M. Nolliastd et al., “Personalizing asr for dysarthric and accented speech with limited data,” arXiv preprint arXiv:1907.13511, 2019.

[6] K. C. Sim, P. Zadrzal, and F. Beaufays, “An investigation into on-device personalization of end-to-end automatic speech recognition models,” arXiv preprint arXiv:1909.06678, 2019.

[7] A. Awasthi, A. Kansal, S. Sarawagi, and P. Jyothi, “Error-driven fixed-budget asr personalization for accented speakers,” in 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7033–7037.

[8] R. Iyer, N. Kharoonganiker, J. Bilmes, and H. Asanani, “Submodular combinatorial information measures with applications in machine learning,” in Algorithmic Learning Theory. PMLR, 2021, pp. 722–754.

[9] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, “Using document summarization techniques for speech data subset selection,” in NAACL, 2013, pp. 721–726.

[10] ——, “Unsupervised submodular subset selection for speech data,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4107–4111.

[11] Y. Wu, R. Zhang, and A. Rudnicky, “Data selection for speech recognition,” in 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU). IEEE, 2007, pp. 562–565.

[12] T. Asami, R. Masumura, H. Masataki, M. Okamoto, and S. Sakauchi, “Training data selection for acoustic modeling via submodular optimization of joint kullback-leibler divergence,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[13] Y. Liu, R. Iyer, K. Kirchhoff, and J. Bilmes, “Switchboard ii and fiswer ii: High-quality limited-complexity corpora of conversational english speech,” in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[14] K. Kirchhoff and J. Bilmes, “Submodularity for data selection in machine translation,” in EMNLP, 2014.

[15] Y. Yuan, S.-W. Chang, and H.-G. Kang, “Gradient-based active learning query strategy for end-to-end speech recognition,” in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 2832–2836.

[16] A. R. Syed, A. Rosenberg, and M. Mandel, “Active learning for low-resource speech recognition: Impact of selection size and language modeling data,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5315–5319.

[17] Y. Hamanaka, K. Shinoda, S. Furui, T. Emori, and T. Koshinaka, “Speech modeling based on committee-based active learning,” in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2010, pp. 4350–4353.

[18] K. Tomanek, F. Beaufays, J. Cattiau, A. Chandorkar, and K. C. Sim, “On-device personalization of automatic speech recognition models for disordered speech,” arXiv preprint arXiv:2106.10259, 2021.

[19] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, “Domain adversarial training for accented speech recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4854–4858.

[20] V. Kaushal, S. Kothawade, G. Ramakrishnan, J. Bilmes, H. Asanani, and R. Iyer, “A unified framework for generic, query-focused, privacy preserving and update summarization using submodular information measures,” arXiv preprint arXiv:2010.05631, 2020.

[21] V. Kaushal, S. Kothawade, G. Ramakrishnan, J. Bilmes, and R. Iyer, “Prism: A unified framework of parameterized submodular information measures for targeted data subset selection and summarization,” in ACM-SIAM Symposium on Discrete Algorithms, 2020.

[22] S. Kothawade, N. Beck, K. Killamsetty, and R. Iyer, “Similar: Submodular information measures based active learning in realistic scenarios,” Advances in Neural Information Processing Systems, vol. 34, 2021.

[23] S. Fujishige, Submodular functions and optimization. Elsevier, 2005.

[24] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, “An analysis of approximations for maximizing submodular set functions—I,” Mathematical programming, vol. 14, no. 1, pp. 265–294, 1978.

[25] A. Gupta and R. Levin, “The online submodular cover problem,” in ACM-SIAM Symposium on Discrete Algorithms, 2020.

[26] H. Lin and J. Bilmes, “Multi-document summarization via budgeted maximization of submodular functions,” in NAACL-HLT, 2010, pp. 912–920.

[27] S. R. Vignesh, S. A. Shanmugam, and H. A. Murthy, “Significance of pseudo-syllables in building better acoustic models for indian english tts,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5620–5624.

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Moyer, R. Morais, L. Saunders, A. Chandorkar, and K. C. Sim, “On-device personalization of automatic speech recognition,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 6124–6128.

[29] 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). IEEE, 2015, pp. 5206–5210.

[30] 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, 2006, pp. 369–376.

[31] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with lightweight convolutions,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4854–4858.

[32] R. Iyer, N. Kharoonganiker, J. Bilmes, and H. Asanani, “Common voice: A massively-multiplangual speech corpus,” arXiv preprint arXiv:1912.06670, 2019.

[33] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” 2020.

[34] S. C. Krotov, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartet: Deep automatic speech recognition with low-resource changes,” in Proceedings of the 23rd international conference on Machine learning, 2006, pp. 369–376.

[35] S. C. Krotov, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, H. Nguyen, Y. Zhang, and J. M. Cohen, “Stochastic gradient methods with layer-wise adaptive moments for training of deep networks,” arXiv preprint arXiv:1905.11286, 2019.