A Toolbox for Construction and Analysis of Speech Datasets: Supplementary Materials

Evelina Bakhturina **Vitaly Lavrukhin** **Boris Ginsburg**

NVIDIA, Santa Clara, USA
{ebakhturina, vlavrukhin, bginsburg}@nvidia.com

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The CTC-Segmentation requires audio and corresponding text along with a pre-trained CTC-based ASR model, see Section 2.1. Limitations of SDE are discussed in Section 2.2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] The Conclusion discusses the need of taking data bias into account and strongly suggest to seek explicit consent from voice actors to use their voices for synthetic replication.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Section 3.2. describes how to fine-tune the Russian QuartzNet model and the data used (MCV ver. 5.1 and RuLS).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.2., the hyperparameters were chosen based on [25]

35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] The experiments were conducted once as the purpose of the paper is to showcase the proposed tools.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.2.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] CTC-Segmentation algorithm is cited in Section 2.1. The datasets used for model fine-tuning and data exploration are cited in Figure 1. Both NeMo code and models are cited accordingly.

(b) Did you mention the license of the assets? [Yes] Both NeMo framework and CTC-Segmentation released under Apache-2.0 License, see Section 2.1. Dataset licenses are mentioned in Figure 1. The RuLS license is mentioned in Section 3.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Both the code for NeMo CTC-Segmentation and SDE tools are released in NeMo framework `https://github.com/NVIDIA/NeMo`

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The datasets used to demonstrate the tools for speech dataset construction and analysis have permissive licenses see Section 3. and Figure 1.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 3.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

The paper focuses on tools and best practices for speech dataset creation and analysis. We demonstrate how to build speech datasets from long audio recordings with loose transcripts using CTC segmentation tool on Russian LibriVox audiobooks. And we describe best practices on how to analyze speech datasets using Speech Data Explorer on MLS and MCV. The developed tools are open-sourced under Apache-2.0 license in NeMo framework `https://github.com/NVIDIA/NeMo`.

Relevant links:

- The documentation for the toolkit could be found at `https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/tools/intro.html`.
- The code is located at `https://github.com/NVIDIA/NeMo/tree/main/tools`.
- A CTC-Segmentation tutorial demonstrates an example on how to cut a LibriVox audio file, see `https://colab.research.google.com/github/NVIDIA/NeMo/blob/stable/tutorials/tools/CTC_Segmentation_Tutorial.ipynb`.
- Fine-tuning of a pre-trained English ASR model on a target language audio data is shown in `https://colab.research.google.com/github/NVIDIA/NeMo/blob/stable/tutorials/asr/ASR_CTC_Language_Finetuning.ipynb`.
- To demonstrate both the CTC-Segmentation and Speech Data Explorer tools, we concatenated the audio files from the LibriSpeech dev-clean split [32] into a single file and set up the toolkit to cut the long audio file into original utterances. We used QuartzNet15x5Base-En model[1] for both segmenation and evaluation. The segmented corpus contains 300 out of the initial 323 minutes of audio. The remaining 23 minutes is the silence removed from the audio's beginning and end during the segmentation. A demo instance of the SDE displays the re-segmented corpus `https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/main/`

[1] `https://ngc.nvidia.com/catalog/models/nvidia:nemospeechmodels`

`tools/speech_data_explorer.html#sde-demo-instance`. Misalignment errors in the segmentation result in a larger gap between ASR model predictions and the reference text. The WER of the re-segmented dataset is 3.82% WER, and CER is 1.16%. Overall, WER and CER of the original LibriSpeech corpus are 3.78% and 1.14%. The slight difference in the scores between the original and the re-segmented corpus can be explained by different model's normalization coefficients due to slightly different segments (as the segmentation dropped some non-speech parts). We manually inspected the examples with the highest WER (CER) values, and the corresponding segmented audio clips sound correct.

For papers introducing best practices in creating or curating datasets and benchmarks, the supplementary materials are not required.