CrowdSpeech and Vox DIY: Benchmark Datasets for Crowdsourced Audio Transcription

Nikita Pavlichenko  
Yandex  
Moscow, Russia  
pavlichenko@yandex-team.ru

Ivan Stelmakh  
Carnegie Mellon University  
Pittsburgh, PA, USA  
stiv@cs.cmu.edu

Dmitry Ustalov  
Yandex  
Saint Petersburg, Russia  
dustalov@yandex-team.ru

Abstract

Domain-specific data is the crux of the successful transfer of machine learning systems from benchmarks to real life. Crowdsourcing has become one of the standard tools for cheap and time-efficient data collection for simple problems such as image classification: thanks in large part to advances in research on aggregation methods. However, the applicability of crowdsourcing to more complex tasks (e.g., speech recognition) remains limited due to the lack of principled aggregation methods for these modalities. The main obstacle towards designing advanced aggregation methods is the absence of training data, and in this work, we focus on bridging this gap in speech recognition. For this, we collect and release CrowdSpeech — the first publicly available large-scale dataset of crowdsourced audio transcriptions. Evaluation of existing aggregation methods on our data shows room for improvement, suggesting that our work may entail the design of better algorithms. At a higher level, we also contribute to the more general challenge of collecting high-quality datasets using crowdsourcing: we develop a principled pipeline for constructing datasets of crowdsourced audio transcriptions in any novel domain. We show its applicability on an under-resourced language by constructing VoxDIY — a counterpart of CrowdSpeech for the Russian language. We also release the code that allows a full replication of our data collection pipeline and share various insights on best practices of data collection via crowdsourcing.

1 Introduction

Speech recognition is an important research problem that has found its applications in various areas from voice assistants such as Siri or Alexa [16] to call centers [32] and accessibility tools [41]. The research community has been actively developing tools for automated speech recognition [11, 26, 35, 40, 46, 21, 36, 47, 22, 15 and many other works]. As a result, the state-of-the-art methods achieve near-perfect performance [49] on LibriSpeech [53] — a famous benchmark to compare speech recognition systems.

[https://github.com/pilot7747/VoxDIY](https://github.com/pilot7747/VoxDIY) Our code is available under the Apache license 2.0, and datasets are available under the CC BY 4.0 license. After the review of this paper, the data will be uploaded to the Zenodo repository and will be provided with a DOI.

While the technical performance on curated benchmarks is almost perfect, it does not necessarily result in reliable, practical performance [41]. Indeed, in real applications, people may use some specific vocabulary or dialects underrepresented in the conventional training data. Thus, blind application of methods trained on the standard benchmarks may result in low accuracy or, perhaps more concerning, discrimination of some subpopulations. For example, a recent study of YouTube’s Automatic Captions reveals a difference in accuracy across gender and dialect of the speaker [42].

One approach towards improving the practical performance of speech-recognition systems is to fine-tune the models in these systems on domain-specific ground truth data. Fine-tuning is very important and efficient for the speech-recognition task [8, 49], but the main problem with this approach is the lack of data. Indeed, even datasets that are considered to be very small in the area (e.g., CHiME-6 [44]) contain hours of annotated audios. While getting such an amount of unlabeled data in a speech-intensive application may be feasible, annotating this data with the help of expert annotators may be prohibitively expensive or slow.

Recently, crowdsourcing has become an appealing alternative to the conventional way of labeling data by a small number of experts. Platforms like Mechanical Turk (https://www.mturk.com/) and Toloka (https://toloka.ai/) significantly reduce the time and cost of data labeling by providing on-demand access to a large crowd of workers. Of course, this flexibility comes at some expense, and the main challenge with the crowdsourcing paradigm is that individual workers are noisy and may produce low-quality results. A long line of work [12, 7, 38, 59, 55, 50] and others] has designed various methods to estimate true answers from noisy workers’ responses to address this issue in multiclass classification. As a result, crowdsourcing has become an industry standard for image labeling. Small and large technology companies continuously use it to improve their services.

In speech recognition, however, the annotations obtained from crowd workers are sentences and not discrete labels, which makes the aforementioned classification methods impractical. Unfortunately, the problem of learning from noisy textual responses and other non-conventional modalities is much less studied in Machine Learning and Computer Science communities. The main obstacle towards solving this problem in a principled manner is the lack of training data: in contrast to the classification setup, worker answers in the speech recognition tasks are high-dimensional, and researchers need a large amount of data to build and evaluate new methods. Therefore, we focus our work on bridging this gap by constructing and analyzing the large-scale dataset of crowdsourced audio transcriptions.

At a higher level, this work also considers the more general challenge of collecting high-quality datasets for under-resourced domains and applications. In many areas, data is the key resource for research. When the costs of getting data are high, it becomes available to only a privileged population of researchers, contributing to the overall inequity of academia. Crowdsourcing offers an appealing opportunity to make science more equitable by making data collection affordable. However, to take the full benefits of crowdsourcing, the research community needs to develop procedures and practices to take reliable control over the quality of collected data. In this work, we build on our long experience of industrial data collection at Yandex and share resources as well as insights that may benefit researchers and engineers who want to collect reliable data on crowdsourcing platforms.

**Our Contributions**

Overall, in this work we make several contributions:

First, we collect and release CROWDSPEECH — the first publicly available large-scale dataset of crowdsourced audio annotations. In that, we obtain annotations for more than 20 hours of English speech from more than 1,000 crowd workers.

Second, we propose a fully automated pipeline to construct semi-synthetic datasets of crowdsourced audio annotations in under-resourced domains. Using this procedure, we construct VoxDIY — a counterpart of CROWDSPEECH for Russian language.

Third, we evaluate the performance of several baseline methods for aggregation of noisy transcriptions on CROWDSPEECH and VoxDIY. Our comparisons indicate room for improvement, suggesting that our data may entail progress in designing better algorithms for crowdsourcing speech annotation.

Fourth and finally, we release the code to fully replicate the data preparation and data collection processes we execute in this work. Additionally, we share various actionable insights that researchers and practitioners can use to fulfill their data collection needs.

The remainder of this paper is organized as follows. We begin from a survey of related work in Section 2. We then construct a pool of speech recordings for annotation in Section 3 and describe the
While avoiding automated aggregation of texts. One approach is to reduce the problem to the conventional classification problem.

Another relevant contribution is a small dataset of translations from Japanese to English constructed by Li. Each sentence in this dataset is associated with ten crowdsourced translations and a ground truth translation. We treat this data as a baseline dataset and return to it in Sections 5 and 6.

Several other works construct benchmark datasets for automated speech recognition without relying on crowdsourced annotation of audios. Specifically, LIBRISPEECH builds on audios with known transcriptions (e.g., audio books or videos with human-generated captions). Starting from annotated long audios, they split the recordings into smaller segments and carefully align these segments with the ground truth texts. While this approach may result in high fidelity datasets, its applicability is limited to domains with pre-existing annotated recordings. Another clever approach is used in the COMMONVOICE dataset. COMMONVOICE is constructed by starting from short ground truth texts and then crowdsourcing speech recordings of these texts. We note that this approach is complementary to ours (start from audios and then crowd-source transcriptions) and the choice between the two approaches may be application-dependent.

Other Approaches. Several papers develop procedures to crowdsource high-quality annotations while avoiding automated aggregation of texts. One approach is to develop a multi-stage process in which workers improve initial transcriptions in subsequent stages, eventually producing a single final transcription. While these approaches offer an appealing alternative to automated aggregation, such pipelines are typically much more complicated and may provide unstable quality due to variations in workers’ accuracy. Another approach is to reduce the aggregation of textual annotations to the conventional multiclass classification. Specifically, given noisy annotations obtained in the first stage, a requester can hire an additional pool of workers to listen to original recordings and vote for the best annotation from the given set. This approach avoids the challenging step of learning from noisy texts by reducing the problem to the conventional classification problem.

However, it is associated with increased costs, and the accuracy of this approach is fundamentally...
Table 1: Summary statistics for the source data used in this work. “Spkrs” stands for “speakers”, letters M and F stand for male and female, respectively.

<table>
<thead>
<tr>
<th>Source Dataset</th>
<th>Version</th>
<th>Language</th>
<th>Nature</th>
<th>Length, hrs</th>
<th># Recordings</th>
<th># F Spkrs</th>
<th># M Spkrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIBRISPEECH</td>
<td>dev-clean</td>
<td>English</td>
<td>Real</td>
<td>5.4</td>
<td>2703</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>test-clean</td>
<td></td>
<td></td>
<td>5.4</td>
<td>2620</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>dev-other</td>
<td></td>
<td></td>
<td>5.3</td>
<td>2864</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td></td>
<td></td>
<td>5.1</td>
<td>2939</td>
<td>17</td>
<td>16</td>
</tr>
</tbody>
</table>

| RUSNEWS        | Ru       | Russian  | Synthetic | 4.8         | 3091         | 1         | 1         |

limited by the accuracy of the best annotation produced in the first stage. In contrast, automated aggregation methods can construct a transcription that is better than all transcriptions in the input.

3 Data Source

In this section, we begin the description of our data collection procedure by introducing the pool of speech recordings that we annotate in Section 4. Table 1 gives an overview of our data sources.

**LibriSpeech Benchmark (cf. Contribution 1)** LibriSpeech is a famous benchmark for comparing speech recognition systems that consists of approximately 1,000 hours of read English speech derived from audiobooks (https://www.openslr.org/12) and split into small segments. An important feature of the dataset is its gender balance – approximately half of the recordings are made by female speakers. In this work, we use the test and development sets of the LibriSpeech benchmark as our initial pool of recordings. Specifically, LibriSpeech consists of two subsets — “clean” and “other”. The clean subset contains recordings of higher quality with accents of the speaker being closer to the US English, while the other subset contains recordings that are more challenging for recognition. To achieve a better diversity of our data, we use both the other and clean subsets of LibriSpeech.

**Under-Resourced Domains (cf. Contribution 2)** LibriSpeech is a rich source of recordings, but it is focused on a specific domain of audiobooks and contains recordings of English speech only. Thus, the aggregation algorithms trained on the annotations of LibriSpeech we collect in Section 4 may not generalize to other domains and languages due to potential differences in workers’ behaviour. To alleviate this issue, we propose the following pipeline to obtain domain-specific datasets for fine-tuning of aggregation methods:

1. Obtain texts from the target under-resourced domain.
2. Use speech synthesis tools to construct recordings of texts collected in the previous step.3
3. Obtain annotations of these recordings using crowdsourcing.

The dataset constructed in this procedure can be used to fine-tune aggregation methods on data from novel domains and languages. In this work, we demonstrate this pipeline by collecting a dataset of crowdsourced annotations of Russian speech recordings. For this, let us now describe the first two steps of the pipeline and introduce the second pool of synthetic recordings that we will annotate in Section 4.

**Texts from a Target Domain** For the sake of the example, we use the domain of news as our target domain. For this, we take sentences in Russian from the test set of the machine translation shared task executed as a part of the Eights and Ninth Workshop on Statistical Machine Translation [6, 7]. To support reliable evaluation, we additionally filter these texts to align their formatting with that used in the LibriSpeech dataset. A detailed description of this stage is given in Supplementary Materials.

**Recording** Synthetic recording is the crux of our approach as it allows us to obtain recordings with known ground truth transcriptions without involving costly human speakers. In this example, we rely on Yandex SpeechKit — an industry-level tool for speech synthesis — to obtain recordings of the ground truth texts. Importantly, Yandex SpeechKit gives access to both “male” and “female” voices, as well as to different intonations (neutral and evil). Thus, in the recording stage, we choose

Note that this data captures the behavior of real workers in the target domain modulo potential differences induced by the use of a synthetic speech generator.
the “gender” and intonation for each recording uniformly at random, ensuring the diversity of our synthetic dataset.

Following the procedure outlined above, we obtain 3091 recordings of Russian speech that we title RUSNEWS. Table 1 gives summary statistics for RUSNEWS and LIBRISPEECH. In the next section, we will use audios from LIBRISPEECH and RUSNEWS to construct datasets CROWDSPEECH (based on LIBRISPEECH) and VOXDIY (based on RUSNEWS) of crowdsourced audio annotations.

4 Data Annotation

With datasets of speech recordings described and prepared in Section 3, we now proceed to the annotation stage in which we build our CROWDSPEECH and VOXDIY datasets. In this section, we introduce the pipeline to gather reliable transcriptions at scale. Figure 1 displays a schematic representation of our data annotation pipeline. It consists of two main components — the entrance exam and the main task, which are jointly designed to provide high-quality data. In what follows, we discuss this pipeline in detail. Additionally, throughout this section, we reflect on our data collection experience and leave practical comments that may be useful for researchers and practitioners.

4.1 Task Design

We begin the exposition of our pipeline with a discussion of instructions, interface, and compensations.

Instructions and Interface Despite the fact that the task of audio transcription may sound natural to workers, there are several important nuances that should be captured in the instructions and the interface. First, given that our ground truth annotations have some specific formatting, we put a strong emphasis on conveying the transcription rules to the workers in the instructions. Next, throughout the task, workers may experience technical difficulties with some recordings, and we design the interface with that in mind. Specifically, at the beginning of the task, workers are asked whether the given audio plays well in their browser. The positive answer to this question serves as a trigger for the text field, and the negative answer allows workers to indicate the technical issue without contaminating our data with a random or empty transcription. The full version of our instructions and a screenshot of the interface are available in Supplementary Materials.

Compensation and Well-Being The recordings we annotate in this work can roughly be grouped by the level of difficulty: RUSNEWS and the clean subset of LIBRISPEECH consist of relatively easy recordings. In contrast, the other subset of LIBRISPEECH has recordings that are harder to annotate. Thus, we issue a compensation of one cent (respectively, three cents) per completed annotation for recordings in the first (respectively, second) group. This amount of compensation was selected for the following reasons: (i) a typical amount of compensation for similar tasks on the Toloka crowdsourcing platform is one cent per recording; (ii) several past speech recognition studies that employed crowdsourcing [3, 28, 33] were issuing a compensation ranging from 0.5 cents to 5 cents for annotation of a recording of comparable length; (iii) a large fraction of workers on Toloka and other crowdsourcing platforms are residents of developing countries with a low minimum hourly wage.

Throughout the experiment, we were monitoring various quantities related to workers’ well-being. Specifically, the hourly compensation for active workers was close to or even exceeded the minimum
hourly wage in Russia — the country of residence for primary investigators of this study. Additionally, our projects received quality ratings of 4.5 and above (out of 5) in anonymous surveys of workers.

Practical Comments. Let us now make several practical comments on our experience:

• First, in preliminary trials, we experimented with increasing the amount of compensation above the aforementioned levels. However, this intervention resulted in our tasks being attractive to spammers and negatively affected the quality of the data we collect. Thus, we decided to stick to the conventional amount of compensation.

• Second, in preliminary trials, we experimented with issuing compensations to workers even when they were unable to play the audio due to self-reported technical difficulties. Unfortunately, this setup resulted in workers reporting technical difficulties for a huge share of the tasks. Once we switched to the compensation for annotated recordings only, the amount of self-reported technical problems reduced drastically without affecting the quality of annotations. This observation suggests a spamming behavior in the original setup.

4.2 Worker Selection and Quality Control

Another key aspect of crowdsourcing data collection is to recruit the right population of workers for the task. For this, we make our task available to only those workers who self-report the knowledge of the language of the task: English for CrowdSpeech and Russian for VoxDIY. Additionally, we implement an Entrance Exam. For this, we ask all incoming eligible workers to annotate ten audio recordings. We then compute our target metric — Word Error Rate (WER) — on these recordings and accept to the main task all workers who achieve WER of 40% or less (the smaller the value of the metric, the higher the quality of annotation). In total, the acceptance rate of our exam was 64%.

Importantly, to achieve a consistent level of annotation quality, it is crucial to control the ability of workers not only at the beginning but also throughout the task. For this, we implement the following rules to detect spammers and workers who consistently provide erroneous annotations:

• Spam-detection rules. Spammers often try to complete as many tasks as possible before getting detected and removed from the platform. To mitigate this behavior, we use a rule that automatically blocks workers from our projects if they complete two or more tasks in less than ten seconds.

• Golden set questions. We use golden set questions to continuously monitor the quality of annotations supplied by workers. If the mean value of the WER metric over the last five completed golden set questions was reaching 35%, we were blocking the worker from taking more tasks.  

Practical Comment. Workers may be hesitant to participate in the tasks if there is a risk that all their work is rejected without compensation. To avoid the additional burden on workers, we follow best practices of (i) compensating the exam for all workers who attempted it, irrespective of whether they passed the bar for the main task or not; (ii) issuing compensations to the workers for the tasks they have completed before being flagged by the our quality control rules.

4.3 Running the Task

Having the pipeline prepared, we annotate each of the five sets of recordings described in Table 1 to construct our CrowdSpeech and VoxDIY datasets. Specifically, we annotate each set of recordings in a separate pool, keeping the task setup identical modulo the use of instructions in Russian for the VoxDIY dataset. Each individual recording was annotated by seven workers, and if any of the assigned workers reported technical difficulties, we reassigned their task to another worker.

As a result of this procedure, we obtained five pools of annotated recordings — first four of these pools comprise the CrowdSpeech dataset and the last pool of Russian recordings comprises the VoxDIY dataset. We release annotated data as well as the Python code to replicate our pipeline in the GitHub repository referenced on the first page of this manuscript.

Privacy Remark The Toloka crowdsourcing platform associates workers with unique identifiers and returns these identifiers to the requester. To further protect the data, we additionally encode each identifier with an integer that is eventually reported in our released datasets.

In this work, we treat all recordings as golden set questions. In practice, one could annotate a handful of examples manually and use them as golden set questions.
5 Exploratory Analysis of Collected Datasets

Having constructed the CROWDSPEECH and VOXDIY datasets, we now proceed to the analysis of collected data on various dimensions, specifically focusing on the reliability of annotators. Before we delve into details, let us make some important remarks.

First, for the sake of analysis, we slightly post-process the annotations obtained in our datasets by removing punctuation marks and making all sentences lowercased. This post-processing step is only needed to ensure consistency with the ground truth data but does not conceptually affect the quality of collected data. Second, when possible, we compare our datasets with CROWDCSA2019—a dataset of crowdsourced translations (with a ground truth translation) constructed by Li et al. [25]. While this dataset is constructed in a different application, it is the largest publicly available dataset for the problem of noisy text aggregation. Hence, it is interesting to juxtapose it to our data. With these preliminaries, we are now ready to present the exploratory analysis of collected datasets.

5.1 Overview of Annotated Datasets

A general overview of the collected datasets is presented in Table 2. First, observe that in total, we have collected 99,519 annotations of 11,126 recordings made by 3,221 unique workers. Thus, our datasets are several orders of magnitude larger than CROWDCSA2019. To the best of our knowledge, our data is also the largest publicly available data of crowdsourced texts.

Second, it appears that in all datasets, the mean length of the crowdsourced annotations (translations) is slightly smaller than the mean length of the ground truth texts. This observation suggests that workers tend to skip some words in both the annotation and translation tasks.

Finally, Figure 2 shows the distribution of the number of tasks completed by a worker for data collected in this study. Observe that these distributions differ significantly between projects, likely being dependent on the difficulty of the task. That would be interesting to see if the aggregation algorithms can adapt for the changing distribution to provide a consistent improvement on different kinds of projects.

Table 2: Overview of datasets collected in this work.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version</th>
<th>Sentence Length, words</th>
<th>Ground Truth</th>
<th>Crowdsourced</th>
<th># Recordings</th>
<th># Workers</th>
<th># Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROWDSPEECH</td>
<td>dev-clean</td>
<td>20.1</td>
<td>19.5</td>
<td>2,703</td>
<td>748</td>
<td>18,921</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dev-other</td>
<td>17.8</td>
<td>16.8</td>
<td>2,864</td>
<td>1,353</td>
<td>20,048</td>
<td></td>
</tr>
<tr>
<td></td>
<td>test-clean</td>
<td>20.1</td>
<td>19.2</td>
<td>2,620</td>
<td>769</td>
<td>18,340</td>
<td></td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td>17.8</td>
<td>16.8</td>
<td>2,939</td>
<td>1,441</td>
<td>20,573</td>
<td></td>
</tr>
<tr>
<td>VOXDIY</td>
<td>RU</td>
<td>13.8</td>
<td>13.6</td>
<td>3091</td>
<td>457</td>
<td>21,637</td>
<td></td>
</tr>
<tr>
<td>CROWDCSA2019</td>
<td>J1</td>
<td>9.5</td>
<td>9.3</td>
<td>250</td>
<td>70</td>
<td>2,490</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>11.9</td>
<td>9.1</td>
<td>100</td>
<td>42</td>
<td>1,000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>11.8</td>
<td>8.6</td>
<td>100</td>
<td>43</td>
<td>1,000</td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Inter-rater agreement according to the Krippendorff’s $\alpha$ with Levenshtein distance. Higher values indicate higher reliability.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version</th>
<th>Overlap</th>
<th>Krippendorff’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROWDSPEECH</td>
<td>dev-clean</td>
<td>7</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>dev-other</td>
<td>7</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>test-clean</td>
<td>7</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td>7</td>
<td>0.78</td>
</tr>
<tr>
<td>VOXDIY</td>
<td>RU</td>
<td>7</td>
<td>0.96</td>
</tr>
<tr>
<td>CROWDCSA2019</td>
<td>T1</td>
<td>10</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>10</td>
<td>0.42</td>
</tr>
</tbody>
</table>

5.2 Inter-Rater Agreement

To evaluate whether the workers understood our task correctly, we compute Krippendorff’s $\alpha$ [19], a chance-corrected inter-rater agreement measure that handles missing values and allows an arbitrary distance function. In this work, we compute the value of Krippendorff’s $\alpha$ using the Levenshtein distance (i.e., edit distance). Since the computation of $\alpha$ is time-consuming as it iterates over all the possible co-occurring item pairs, we obtain and report the sampling estimate of this value as follows. For each sample in the set of 10,000 samples, we randomly select 100 different audio recordings with replacement and compute $\alpha$ for all the transcriptions obtained for these recordings. We then take the mean of these values across all iterations and report it in Table 3.

When interpreting the results, we follow the recommendations of Krippendorff [19]. The values of $\alpha \geq 0.8$ suggest that annotations we obtained for our dataset are reliable. Thus, we conclude that workers on Toloka have successfully understood and performed the given audio transcription task.

Interestingly, the CROWDCSA2019 dataset demonstrates much lower agreement between raters. We hypothesize that the discrepancy between this dataset and our datasets is due to the different natures of the tasks. Indeed, translation is a much more subjective task as compared to speech recognition — in the case of translations, there could be multiple equally good translations, and even ideal translators may have some disagreement. In contrast, the audio transcription task has a unique underlying ground truth transcription, and hence ideal annotators can achieve a perfect agreement.

6 Evaluation

In this section, we evaluate the existing methods for aggregation of noisy texts on our datasets.

6.1 Methods

In our evaluations, we consider the following baseline methods, using implementations from the Crowd-Kit library (https://github.com/Toloka/crowd-kit) when available.

- **Random** It is a naive baseline that always chooses one of the annotations provided for the task uniformly at random.

- **ROVER** Recognizer Output Voting Error Reduction [14] is our first non-trivial baseline. The method was originally designed to combine the output of several different automatic speech recognition systems but was also demonstrated to perform well on crowdsourced sequences [3, 13, 24, 25]. Under the hood, it performs an alignment of given sequences using dynamic programming and then computes the majority vote on each token.

- **RASA** Reliability Aware Sequence Aggregation [25] employs large-scale language models for aggregating texts. It encodes all worker responses using RoBERTa [27] (RuBERT [20] for the Russian language) and iteratively updates the mean weighted embedding of workers’ answers together with estimates of workers’ reliabilities. Finally, the method defines the final answer to be the response closest to the aggregated embedding based on the notion of cosine distance.
In parallel with preparing this paper, we have designed and executed a shared task (https://crowdscience.ai/challenges/vldb21) on developing aggregation methods for crowdsourced audio transcriptions. The competitive nature of the task does not allow us to use LIBRI SPEECH audios.

To perform the qualitative evaluation of the baseline methods, we execute them on each of the datasets and report it in Table 4. First, observe that on the CROWDSPEECH dataset, there is a consistent gap between all baselines and the oracle performance, with the gap being larger for more difficult subsets (dev-other and test-other). This observation indicates that there is room for the development of better aggregation methods that keep up with, or even exceed, the performance of the Oracle.

Second, we note that when the quality of recordings is good, as in the case of the semi-synthetic VoxDIY dataset, baseline methods achieve a near-perfect performance. This observation suggests that for putting aggregation methods in a more realistic (challenging) scenario, one can additionally corrupt synthetic recordings with background noise.

Finally, we note that the performance of all aggregation methods, including Oracle, is much weaker on the CROWDCSA2019 dataset. This effect is likely an artifact of the subjective nature of the machine translation task, which, in contrast to the speech recognition task, does not have a unique ground truth answer. Thus, CROWDCSA2019 may not be the best choice to design aggregation methods for objective tasks such as speech recognition. The same observation applies to the methods trained on this dataset: Table 4 indicates that a simple ROVER baseline is always superior to more advanced algorithms developed for CROWDCSA2019. Of course, a symmetric observation applies to our CROWDSPEECH and VoxDIY which may not be the best choice to design aggregation methods.

### Table 4: Comparison of the baselines and the oracle performance. Evaluation criterion is the average word error rate (WER) and lower values are better.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version</th>
<th>Random</th>
<th>Oracle</th>
<th>ROVER</th>
<th>RASA</th>
<th>HRRASA</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROWDSPEECH</td>
<td>dev-clean</td>
<td>17.39</td>
<td>3.81</td>
<td>6.76</td>
<td>7.50</td>
<td>7.45</td>
</tr>
<tr>
<td></td>
<td>dev-other</td>
<td>27.73</td>
<td>8.26</td>
<td>13.19</td>
<td>14.21</td>
<td>14.20</td>
</tr>
<tr>
<td></td>
<td>test-clean</td>
<td>18.89</td>
<td>4.32</td>
<td>7.29</td>
<td>8.60</td>
<td>8.59</td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td>27.28</td>
<td>8.50</td>
<td>13.41</td>
<td>15.67</td>
<td>15.66</td>
</tr>
<tr>
<td>VoxDIY</td>
<td>RU</td>
<td>7.09</td>
<td>0.70</td>
<td>1.92</td>
<td>2.22</td>
<td>2.20</td>
</tr>
<tr>
<td>CROWD CSA2019</td>
<td>J1</td>
<td>76.64</td>
<td>36.50</td>
<td>61.16</td>
<td>65.86</td>
<td>67.57</td>
</tr>
<tr>
<td></td>
<td>T1</td>
<td>63.08</td>
<td>28.07</td>
<td>51.35</td>
<td>48.29</td>
<td>49.99</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td>63.69</td>
<td>30.46</td>
<td>52.44</td>
<td>49.82</td>
<td>52.04</td>
</tr>
</tbody>
</table>

- **HRRASA** We also use a modification of RASA called HRRASA [24] that, besides the global reliabilities, uses local reliabilities represented by the distance from a particular response to other responses for the task. In the original paper [24], GLEU [20, 31] metric was suggested to calculate the distance between sequences, and we resort to this choice in our experiments as well.

In addition to comparing the baselines, we want to make a rough conclusion on whether any of them demonstrate the optimal performance on our data. Note that it may be infeasible to uncover all transcriptions with absolute accuracy as for some recordings, the noise in annotations could vanish out all the signal. To obtain a more reasonable estimate of achievable performance, we introduce the **Oracle** aggregation algorithm to the comparison. For each recording, it enjoys the knowledge of the ground truth and selects the best transcription provided by the workers as its answer.

The Oracle method achieves the maximum accuracy that can be reached by an aggregation algorithm restricted to the set of transcriptions provided by the workers. Thus, Oracle gives a weak estimate of the achievable quality as its accuracy could be improved by an algorithm that is allowed to modify transcriptions provided by the workers. Nevertheless, in the analysis below, we focus on the gap between the baselines and the Oracle to estimate if there is some room for improvement on our data.

### 6.2 Results

To perform the qualitative evaluation of the baseline methods, we execute them on each of the datasets under consideration. We then compute the mean value of WER over all recordings in each dataset and report it in Table 4. First, observe that on the CROWDSPEECH dataset, there is a consistent gap between all baselines and the oracle performance, with the gap being larger for more difficult subsets (dev-other and test-other). This observation indicates that there is room for the development of better aggregation methods that keep up with, or even exceed, the performance of the Oracle.

Second, we note that when the quality of recordings is good, as in the case of the semi-synthetic VoxDIY dataset, baseline methods achieve a near-perfect performance. This observation suggests that for putting aggregation methods in a more realistic (challenging) scenario, one can additionally corrupt synthetic recordings with background noise.

Finally, we note that the performance of all aggregation methods, including Oracle, is much weaker on the CROWDCSA2019 dataset. This effect is likely an artifact of the subjective nature of the machine translation task, which, in contrast to the speech recognition task, does not have a unique ground truth answer. Thus, CROWDCSA2019 may not be the best choice to design aggregation methods for objective tasks such as speech recognition. The same observation applies to the methods trained on this dataset: Table 4 indicates that a simple ROVER baseline is always superior to more advanced algorithms developed for CROWDCSA2019. Of course, a symmetric observation applies to our CROWDSPEECH and VoxDIY which may also be suboptimal for the machine translation task.

### 6.3 Methods Developed on Our Data

In parallel with preparing this paper, we have designed and executed a shared task (https://crowdscience.ai/challenges/vldb21) on developing aggregation methods for crowdsourced audio transcriptions. The competitive nature of the task does not allow us to use LIBRI SPEECH audios.
We demonstrated this pipeline by constructing the VoxDIY dataset in the present paper. Specifically, we crowdsourced annotations of synthetic recordings of passages from Wikipedia and books [13].

One of the best results in this shared task was demonstrated by a carefully fine-tuned T5 model [37]. With permission of the author, we evaluate their model on the test sets of the CROWDSPEECH dataset collected in this paper. For this, we introduce three additional models (see details in Appendix D):

- T5 model trained on the shared task data [T5 (ST)]
- T5 (ST) model additionally fine-tuned on the development sets of CROWDSPEECH [T5 (ST+FT)]
- Finally, we annotate additional 11,000 examples from LibriSpeech train-clean-100 subset following the same methodology as described in Section 4. We then fine-tune the original T5 model [37] on this data only [T5 (FT)].

We juxtapose these models to the best of the available baselines (ROVER) on the test sets. Table 5 displays the result of comparison. First, we note that all T5-based models significantly outperform ROVER — the baseline that remained unchallenged in more than twenty years — demonstrating the state-of-the-art result on both test-clean and test-other sets. Additionally, we note the effect of fine-tuning on the domain-specific data, resulting in T5 (ST+FT) model being superior to T5 (ST) on both test sets. Similarly, T5 (FT) outperforms other models on the test-clean set (we trained it on data from the clean subset of LibriSpeech) but demonstrates lower accuracy on the test-other set (it did not get to see data from the other subset of LibriSpeech).

With these observation, we make the following conclusions:

- First, there is initial evidence that the data we release in this paper is instrumental in developing novel principled methods for aggregation of crowdsourced annotations.
- Second, note that the T5-based method was designed on the data collected through the semi-synthetic procedure used to construct VoxDIY. Given that the strong performance of the method carried over to the realistic CROWDSPEECH dataset, we conclude that the semi-synthetic data may be useful to quickly explore new domains in which no recordings of human voice exist.

Table 5: Comparison of the T5-based method developed on our shared task with ROVER — the strongest of the existing baselines. Models are compared on WER and Lower values are better.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Version</th>
<th>Oracle</th>
<th>ROVER</th>
<th>T5 (ST)</th>
<th>T5 (ST+FT)</th>
<th>T5 (FT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROWDSPEECH</td>
<td>test-clean</td>
<td>4.32</td>
<td>7.29</td>
<td>6.21</td>
<td>5.32</td>
<td>5.22</td>
</tr>
<tr>
<td></td>
<td>test-other</td>
<td>8.50</td>
<td>13.41</td>
<td>11.80</td>
<td>10.46</td>
<td>11.67</td>
</tr>
</tbody>
</table>

7 Conclusion

In this work, we collected and released CROWDSPEECH — the first publicly available large-scale dataset of crowdsourced audio transcriptions. Based on evaluations of existing baselines on our data, we believe that our work will enable researchers to develop principled algorithms for learning from noisy texts in the crowdsourcing setting. Additionally, we proposed an automated pipeline for collecting semi-synthetic datasets of crowdsourced audio transcriptions in under-resourced domains. We demonstrated this pipeline by constructing VoxDIY — a Russian counterpart of CROWDSPEECH.

In the end, we should mention some limitations of our work. First, we admit that the use of speech synthesis techniques could affect the distribution of errors people make when annotating audios, thereby affecting the generalization ability of aggregation tools trained on VoxDIY. Second, in this work, we annotated our datasets in an industry-level pipeline, which resulted in annotations of high quality. It would be interesting to additionally collect datasets under less stringent quality control rules or for more challenging data to make our data even more diverse in terms of complexity.

Better understanding and addressing these limitations is an interesting direction for future research work. With these caveats, we encourage researchers and practitioners to use our data judiciously and to carefully evaluate all the risks and benefits in their specific application.

---

3 Implementation is available at https://github.com/AlexRey/VLDB2021_workshop_t5
References

    In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of
    PMLR.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer,
    Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A
    Massively-Multilingual Speech Corpus. In *Proceedings of The 12th Language Resources and
    Language Resources Association (ELRA).

    broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. In
    *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2011,
    pages 4980–4983, Prague, Czech Republic, 2011. IEEE.


[5] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman,
    David R. Karger, David Crowell, and Katrina Panovich. Soylent: A Word Processor with a
    Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software
    and Technology*, UIST ’10, pages 313–322, New York, NY, USA, 2010. ACM.


[7] Ondřej Bojar et al., editors. *Proceedings of the Ninth Workshop on Statistical Machine Transla-

    SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural

[9] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan
    Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe,
    Shuaijiang Zhao, Wei Zou, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You,
    and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of
    transcribed audio, 2021.

[10] Otto Chrons and Sami Sundell. Digitalkoot: Making old archives accessible using crowdsourc-

    a German and French heritage corpus. In *Proceedings of the Tenth International Conference
    on Language Resources and Evaluation (LREC’16)*, pages 975–982, Portorož, Slovenia, May
    2016. European Language Resources Association (ELRA).

    Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society, Series C (Applied

    Transcription of Non-Native Speech. In *Proceedings of the NAACL HLT 2010 Workshop
    on Creating Speech and Language Data with Amazon’s Mechanical Turk*, pages 53–56, Los

    Output Voting Error Reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech
    IEEE.


Checklist

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]
   
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   
   (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]
   
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   
   (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)? [N/A]

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]
   
   (b) Did you mention the license of the assets? [Yes]
   
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes]
   
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]

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? [Yes]
   
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes]
   
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes]