My Science Tutor (MyST) Children’s Speech Corpus
Boulder Learning, Inc.

Version: 0.4.0

Corpus Overview

The Children’s Speech Corpus was created as part of the My Science Tutor (MyST) project. We will refer to it as the MyST corpus. It consists of 393 hours of speech collected across 1,371 students from the 3rd, 4th and 5th grades. Students conversed with a virtual science tutor in 8 areas of science, resulting in a total of 10,496sessions and a total of 228,874utterances. 45% of the utterances have so far been transcribed at the word level.

Students	Sessions	Total Utterances	Transcribed Utterances
1,371	10,496	228,874 (393 hours)	103,082

Partitioning of Data for Development and Evaluation

For the convenience of the ASR community, we partitioned and structured the corpus upfront into training, development and test sets. These partitions were generated ensuring that they reasonably represent each of the science module in MyST and that each student is present in only one of the three partitions. These three data sets are in three separate directories in the corpus release. (See Corpus Structure, below.)

Data Collection

The following section describes the process that was used to collect this data.

Methodology

The MyST corpus was collected in 2 stages—Phase I and Phase II. In both phases, the content covered is aligned to Full Option Science System (FOSS) modules, which typically last 8 weeks during the school year. FOSS is used by over 1 million children in over 100,000 classrooms in all 50 states in the U.S. FOSS modules are centered on science investigations. There are typically 4 Investigations in a module (e.g., in the Magnetism and Electricity module, the 4 investigations are Magnetism, Serial circuits, Parallel Circuits, and Electromagnetism). Each Investigation has 3 to 4 classroom “investigation parts” where groups of students work together to, for example, build a serial circuit to make a motor run, and record their observations in science notebooks. Shortly after conducting an “investigation part”, students interact with a virtual tutor for 15-20 minutes. The tutor asks the student questions about science presented in illustrations, animations or interactive simulations, with follow-up questions designed to stimulate reasoning and help students construct accurate explanations.

The system is strict turn-taking; the tutor presents information, asks a question and waits for the student to respond. To respond, the student presses the spacebar on the laptop, holds it down while speaking, and releases it when done. Each student turn is recorded as a separate audio file. When transcribed, an utterance level transcript file is created for each audio file. No identifying information is stored with the data, only anonymized codes for schools and students. All students and their parents signed consent forms allowing Boulder Learning to enter and distribute their anonymous speech data.

Descriptive Statistics

Some characteristics of the data collected in the two phases is described below.

Phase I

The Phase I corpus contains sessions from students in grades 3-5. All of the sessions from this phase have been transcribed. The following modules were included in this phase.

1. ME - Magnetism and Electricity
2. MS - Mixtures and Solutions
3. VB - Variables
4. WA - Water

```
Number of Students:     421
Number of Sessions:    1509 (102 hours)
Transcribed Sessions:  1509 (102 hours)
Untranscribed Sessions: -
```

During this phase, there was no attempt to have any individual student cover all of the parts for a module. The focus of the collection during this phase was to get a wide variety of students rather than try to get complete coverage of material for individual students.

Phase II

The Phase II corpus contains sessions from students in grades 4-5. It included the following 5 modules, with an average of 10 parts each

1. EE - Energy and Electromagnetism
2. MX - Mixtures
3. SMP - Sun, Moon and Planets
4. SRL - Soil, Rocks and Landforms
5. LS - Living Systems

```
Number of Students:       950
Number of Sessions:     8,987 (291 hours)
Transcribed Sessions:   1,426 ( 95 hours)
Untranscribed Sessions: 3,711 (196 hours)
```

In this collection, teachers were asked to have students complete all parts for 2 modules, however, many teachers did not want to cover 2 modules and whatever data was collected was kept, even if students didn't complete the sequence.

Guidelines

During Phase I of the project we used a rich (slow, expensive) transcription guidelines—the ones typically used by speech recognition researchers. However, we realized that for the purposes of this project, we did not need to get that level of richness in the transcriptions, and therefore during Phase II, we decided to use a reduced (quick, cheaper) version of those guidelines which allowed us to transcribe more data.

Corpus Structure

The directory structure for the corpora is as shown in the figure below. Variables are enclosed in angle-brackets (`<variable>`) and can take values as described immediately after.

```
corpora/
└── myst/
    ├── data/
    |   ├── <partition>/
    |   |    ├── <student_id>/
    |   |    |    └── <session_id>/
    |   |    |         ├── <session_id>.<file-extension>
    |   |    |         ├── ...
    |   |    |         ├── ...
    |   |    |         ├── ...
    ...
    ...
    ...
```

Where,

`<partition>` is one of `train`, `development` or `test`.

`<student_id>` is a 6-digit ID with the first 3 digits representing the school code and the next 3 digits the student number.

`<session_id>` is the ID for a particular session and is further represented as `<corpus>_<student_id>_<date>_<time>_<module>_<investigation>.<part>`

`<date>` is represented as `<YYYY>-<MM>-<DD>`

`<time>` is represented as `<hh>-<mm>-<ss>.` Wherein, `<hh>` represents the hour, `<mm>` represents minute, and `<ss>` represents seconds. In Phase I, we did not capture hour/minute/second for each session, so the corresponding fields for sessions in Phase I are set to `00`

`<module>` is a two- or three-character string enumerated in the Phase sections above.

`<investigation>` is a decimal number representing the respective investigation for a module.

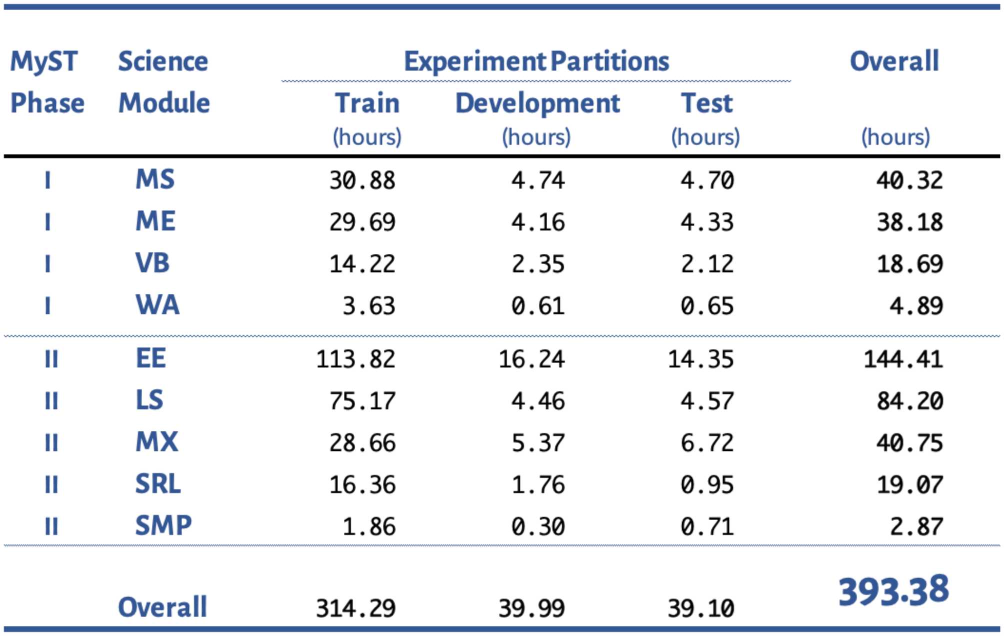
`<part>` is the utterance ID within a session. Numbers `001` onward represent the index of each utterance in a session.

`<file-extension>` is one of the following:

`.wav` – The audio file. Each one representing an utterance.
`.trn` – Transcription of the corresponding audio file

Experiment Partitions

The table below lists the distribution of audio data (in hours of audio) that are present in each of the train, development and test partitions. It shows them aggregated over each partition, each science module[1], and the sum total hours for the entire corpus.



MyST	Science	Experiment Partitions			Overall
Phase	Module	Train	Development	Test	
		(hours)	(hours)	(hours)	(hours)
I	MS	30.88	4.74	4.70	40.32
I	ME	29.69	4.16	4.33	38.18
I	VB	14.22	2.35	2.12	18.69
I	WA	3.63	0.61	0.65	4.89
II	EE	113.82	16.24	14.35	144.41
II	LS	75.17	4.46	4.57	84.20
II	MX	28.66	5.37	6.72	40.75
II	SRL	16.36	1.76	0.95	19.07
II	SMP	1.86	0.30	0.71	2.87
	Overall	314.29	39.99	39.10	393.38

[1] A module is unique to each of the two MyST phases

Data Cleanup and Pre-processing

We did a pass over the corpus to clean up various types of errors that could be identified using statistics on the underlying audio and potentially erroneous data collection.

Ensure Data Provenance

The following checks were carried out on the data before it went into the release.

Consent and Assent

The University of Colorado's Institutional Review Board reviewed and approved all components of the My Science Tutor project to assure student privacy. The review board approved the Parental Consent forms and the Student Assent forms. All utterances in the corpus were signed by a student's parent or guardian, and by the student. The final Parental Consent and Student Assent forms approved by the IRB explicitly provide permission for anonymous student speech data and transcriptions to be distributed for both research and commercial use. We manually verified that we had parental consent and student assent for every student in the corpus.

Session Quality

Bad—empty or corrupted sessions were removed using simple heuristics and based on missing data.

Session Length

Sessions that were less than a certain minimal threshold (< 10 minutes long), or longer than a certain maximum threshold (> 1 hour long) were inspected and corrected or removed.

Missing audio files

Sessions that were missing audio files for a significant number of utterances were deleted.

Audio Quality

All utterances were processed to identify all possible unacceptable recordings and were removed from the database. We performed the following checks for audio quality.

Clipping Rate

If there was a significant number of frames (exceeding a certain threshold) that were clipped, we removed or marked the audio file. We removed them if it impacted more than a certain fraction of utterances in a session. In which case we also removed the session from the release. If only a small number of files had large fraction of clipping, we tagged them in a report file, so that the users can determine whether to include or exclude that data from their study.

Silence

Sometimes there are significant amounts of leading and trailing silence in the audio files. We trimmed all such silence. We did not, however, remove or compress silence that occurred within an utterance.

Background Noise

Utterances with a significant amount of noise or cross talk were removed. This was only possible for the cases that were transcribed or fell in the fraction of sample utterances that we manually verified.

Transcription Quality

We fixed obvious spelling errors in the transcriptions. We tried to retain explicitly mispronounced words as much as possible.

Updated Pronunciation Dictionary

We also make available an updated pronunciation dictionary. We used CMU's pronunciation dictionary as a starting point and added words that were novel to this corpus.