Crowdsourcing tool for annotated speech corpora creation

KHUSAINOV AIDAR
INSTITUTE OF APPLIED SEMIOTICS
KAZAN, RUSSIA
Outline

1. Introduction
2. Broadcast speech annotation
3. Correcting annotations for crowdsourced audio
1. Introduction

«Classical» approach:
• Acoustic models => phonemes
• Pronunciation model => words
• Language model => phrase

End-to-end approaches:
• Better accuracy, require a large amount of training data
• Using data for related languages; pre-trained models

Wav2vec2:
• NLP/Computer Vision fields benefited from using self-supervised pretraining
• Allows to learn robust audio representations based on unlabeled data
• Masking fragments; model tries to distinguish the true speech representation from distractors (uniformly sampled from other masked fragments)
1. Introduction

The main benefits:
- For low-resourced language it’s much easier to find unlabeled data
- ASR system becomes more robust to background noises, dialects, speakers

Main goals:
- Collect required unlabeled and labeled Tatar speech corpora
- Try the approach with iterative self-supervised pretraining steps on audio data that is increasingly closer to the target domain
1. Introduction

- Recordings’ format: 16 kHz, 16 bps mono WAV PCM
- Speakers: native speakers, Kazan dialect
- Speech type: read speech

![Tatar speech corpus diagram]

- "Core" part: 10%
- "Read" part: 84%
- "Spontaneous" part: 6%
1. Introduction

• Core part
  • Manually collected separate words and phrases
  • Phonetically full, max context
  • 251 speaker, average duration – 0:01:58
  • Total duration – 8:12:16

• Read part:
  • Rule-based selection from text corpus
  • 190 speakers, average duration – 0:22:18
  • Total duration – 70:39:00

• Spontaneous part:
  • Non-overlapping dialogues
  • Total duration – 5:19:33
1. Introduction

<table>
<thead>
<tr>
<th>Speech corpus</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># speakers</td>
<td>499</td>
</tr>
<tr>
<td>Duration</td>
<td>99:09:59</td>
</tr>
<tr>
<td>Male / Female</td>
<td>30% / 70%</td>
</tr>
<tr>
<td><em>Spontaneous speech</em></td>
<td>5:19:33</td>
</tr>
</tbody>
</table>

* We’re recording spontaneous speech too, but it’s not annotated
1. Introduction

• **Annotation**
  • Speaker’s name
  • Age
  • Gender
  • Native language
  • Nationality
  • Speech quality (expert’s mark from 1 to 5)
  • Dialect
  • Microphone model
  • Comment
Outline

1. Introduction

2. Broadcast speech annotation

3. Correcting annotations for crowdsourced audio
2. Project description

Main goal – tools for corpus creation.

1. Broadcast speech annotation web-tools
   [self-supervised approaches]

2. Tool to check and correct annotations
2. Project description

- **ASP.Net Core**
- **React.js**
- **DDD (Domain Driven Design):**
  - Infrastructure Layer
  - Domain layer
  - Application Layer
  - Service Layer
  - Presentation Layer
  - Client Applications
2. Project description

PostgreSQL

<table>
<thead>
<tr>
<th>AudioSegments</th>
<th>AudioSegmentAnnotations</th>
<th>AudioSegmentAnnotationValidations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
<td>Id</td>
<td>Id</td>
</tr>
<tr>
<td>ASRText</td>
<td>AudioSegmentId</td>
<td>AudioSegmentAnnotationId</td>
</tr>
<tr>
<td>AnnotationStatus</td>
<td>CreationTime</td>
<td>CreationTime</td>
</tr>
<tr>
<td>AudioFileName</td>
<td>CreatorUserId</td>
<td>CreatorUserId</td>
</tr>
<tr>
<td></td>
<td>Text</td>
<td>IsValid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Id</td>
</tr>
<tr>
<td>EmailAddress</td>
</tr>
<tr>
<td>IsActive</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Password</td>
</tr>
</tbody>
</table>
2. Project description

Basic functionality:

• Audio files upload;
• VAD and splitting uploaded files into fragments;
• Web-form for annotating fragment;
• Web-form for validating made annotations;
• View status of annotation of all segments;
• Downloading the annotations.
2. Project description

View fragments’ statuses
2. Project description

Annotating fragments
2. Project description

Validating fragments
2. Project description

Initial data:
- From TNV Planeta broadcast company;
- Recordings from December 2019;
- AVI video with mp3 96 kB/s stereo audio signal;
- Converted to 16 bps 16 kHz WAV;
- Total duration – 733 hour.
2. Project description

We manually selected segments for the first stage annotation:

- News programs;
- Interviews;
- Talk-shows.

In total 40 segments (23 hours 21 minutes) have been uploaded to the system.

This gave us 22,432 audio fragments with a duration less than 15 seconds.
1. Introduction

2. Broadcast speech annotation

3. Correcting annotations for crowdsourced audio
3. Telegram bot

@TatarVoiceBot

Goal – 500 hours
- 408 speakers
- 9 hours 28 minutes

Commands:
1. Next – new phrase to read and send as voice message;
2. Correct – record previous phrase again;
3. Skip – to skip current phrase;
4. Statistics – show user’s and overall statistics;
5. Age – select age interval;
6. Russian/Tatar/Help.
3. Process

- People read phrases (12 000)
- ASR recognizes text
- Manual check for files with errors (4 200)
3. Data collection

https://github.com/heartexlabs/label-studio:

- Allows to build universal platform for all Institute annotation tasks;
- Easy to configure for each task (interface, data, hotkeys, etc.);
- Local, free.
3. Data collection

Requirements:

• Unlabeled dataset for SS steps;
• Labeled dataset for FT steps.

Labeled datasets:

1. «Tatar Corpus»;
2. «Common Voice»;
3. [new] TV broadcasting;
3. Data collection

Unlabeled dataset:
1. Audiobooks (read speech, recording studio) – 114 hours;
2. TV broadcasting recordings for 1 month (spontaneous speech, bg noises, music) – 733 hours;
3. 2 radio stations archives (read and spontaneous, bg music) – 215 hours;
4. 100 scientific lectures from YouTube (good SNR) – 87 hours.

Preprocessing:
1. Audio track extraction;
2. Audio conversion to 16 bps 16 kHz mono format;
3. VAD;
4. Filtering short (<4.5 sec) and long (>30 sec) fragments.
3. Data collection

Statistics of final unlabeled dataset

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Initial</th>
<th>After VAD</th>
<th>After filtering short and long fragments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Audiobooks</strong></td>
<td>114 hours</td>
<td>105 hours</td>
<td>58 hours</td>
</tr>
<tr>
<td></td>
<td>(520 fragments)</td>
<td>(36 712 fragments)</td>
<td>(17 563 fragments)</td>
</tr>
<tr>
<td>TV</td>
<td>733 hours</td>
<td>472 hours</td>
<td>202 hours</td>
</tr>
<tr>
<td></td>
<td>(62 fragments)</td>
<td>(263 466 fragments)</td>
<td>(67 065 fragments)</td>
</tr>
<tr>
<td><strong>Radio stations</strong></td>
<td>215 hours</td>
<td>146 hours</td>
<td>29 hours</td>
</tr>
<tr>
<td></td>
<td>(398 fragments)</td>
<td>(29 778 fragments)</td>
<td>(8 941 fragments)</td>
</tr>
<tr>
<td><strong>YouTube videos</strong></td>
<td>87 hours</td>
<td>81 hours</td>
<td>39 hours</td>
</tr>
<tr>
<td></td>
<td>(100 fragments)</td>
<td>(31 437 fragments)</td>
<td>(12 764 fragments)</td>
</tr>
<tr>
<td><strong>Corpus</strong></td>
<td>1 150 hours</td>
<td>804 hours</td>
<td>328 hours</td>
</tr>
</tbody>
</table>
Thank you

Khusainov Aidar
khusainov.aidar@gmail.com