Zpoken is a Ukrainian IT company with one of the major divisions oriented on Speech Recognition technologies in English and Slavic (Ukrainian, Russian) languages.

# What is it

We are happy to present here our Russian Speech Dataset - Zpoken Dataset [RU]

At the current moment the dataset consists of 5 source parts: radio\_source\_1, radio\_source\_2, radio\_source\_3, radio\_source\_5, Ru-films.

All data is stored in .opus format and was converted to mono, 16 kHz sampling rate, 16-bit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Part name | Duration  (h) | Samples num. | Average duration (s) | Characters per second | Characters per sample |
| radio\_source\_1 | 16 424.82 | 7 887 042 | 7.50 | 14.12 | 105.84 |
| radio\_source\_2 | 2 308.46 | 955 904 | 8.69 | 13.53 | 117.62 |
| radio\_source\_3 | 500.14 | 165 584 | 10.87 | 13.90 | 151.16 |
| radio\_source\_5 | 655.88 | 216 101 | 10.93 | 16.63 | 181.66 |
| Ru-films | 850.88 | 203 972 | 15.02 | 8.76 | 131.57 |
| Total | Average | **20 740,18** | 9 428 603 | 10.11 | 13.35 | 144.57 |

All parts were scraped from open sources. Basically there were long audio files and transcriptions without timesteps. So that one of the challenges we solved is to align original transcription directly to each short audio sample. More about this problem you can read [here](https://docs.google.com/document/d/1TKXRT-9KFldIuQLsIVIno8Mt2Gjc0JEQfCAm4Lzf9nI/).

# Download and play

We provide **absolutely free to use 150 hours** demos for each part. It is a randomly selected sample from the original dataset part.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Part name | Duration  (h) | Samples num. | Size (MB) | Link to download |
| radio\_source\_1 | 50 | 34 356 | 837 | [Radio1\_50h.zip](https://drive.google.com/file/d/1DiAXSK1DQf2beiN0-aJ3nDPK_f_5-S4C/) |
| radio\_source\_2 | 25 | 16 041 | 430 | [Radio2\_25h.zip](https://drive.google.com/file/d/1GRDrZjjOTfO5WohDX6Q83PgI5VMOyNlm/) |
| radio\_source\_3 | 25 | 8 933 | 418 | [Radio3\_25h.zip](https://drive.google.com/file/d/1FJhT_U4FHZaz5bosEAlfy-77dYn0qir5/) |
| radio\_source\_5 | 25 | 10 786 | 441 | [Radio5\_25h.zip](https://drive.google.com/file/d/1qEsOFMWxFZ4mSvEul2RuaY8JRq075bjZ/) |
| Ru-films | 25 | 7 358 | 380 | [Ru\_films\_25h.zip](https://drive.google.com/file/d/1Ko5WdTvaxME7XP2k6pNjttoc7dgL4I-x/) |
| Total | 150 | 77 474 | 2 506 |  |

# 

# 

# They are hosted on [Gdrive](https://drive.google.com/drive/folders/1PBO_cGd-8wLIaxlmIF7l6XX-CEE7Kl70?usp=sharing) so we provide *./download.sh* to easily get them.

## Requirements

You need the [gdown](https://github.com/wkentaro/gdown) to run the *./download.sh*

*pip install gdown*

Just run *bash download.sh* on your linux machine.

# Data structure

You will find the next directory structure, after you unzip each archive.

+---<DatasetPartName>

| +---data

| | +---subfolder1 (optional)

| | | +---speech\_file1.opus

| | | +...

| | | \---speech\_file[N].opus

| | +...

| | +---subfolder[N] (optional)

| | | +---speech\_file1.opus

| | | ...

| | \ \---speech\_file[N].opus

| +---transcription.csv

# Get full dataset.

If you are interested in the full version of the dataset feel free to contact us in this form. Usually we’ll answer in one working day.

# Future work

* release more hours
* optimize archive storage (Gdrive is too annoying)
* share Jupyter notebook to explore the dataset.

# License

CC-BY-4.0

<HTML>

<a rel="license" href="http://creativecommons.org/licenses/by-sa/4.0/"><img alt="Creative Commons License" style="border-width:0" src="https://i.creativecommons.org/l/by-sa/4.0/88x31.png" /></a><br /><span xmlns:dct="http://purl.org/dc/terms/" href="http://purl.org/dc/dcmitype/Dataset" property="dct:title" rel="dct:type">Zpoken Dataset [RU]</span> is licensed under a <a rel="license" href="http://creativecommons.org/licenses/by-sa/4.0/">Creative Commons Attribution-ShareAlike 4.0 International License</a>.

</HTML>