SNUV

Overview

SNUV (**S**pelling and **NU**mbers **V**oice database) is a Polish speech corpus designed for the development and evaluation of automatic speech recognition (ASR) systems for spoken spelling and number recognition.

The corpus contains **132,656 complete transcribed recordings**, comprising **221.74 hours of speech**, produced by **210 Polish speakers**. The speaker metadata identifies **120 female speakers** and **90 male speakers**. Known speaker ages range from **11 to 69 years**; age information is unavailable for three participants.

The speech material focuses on utterances involving spoken letter names, spelled words, digit sequences and numerical expressions. These are relevant to voice-driven applications in which users communicate names, identifiers, telephone numbers, account numbers, registration numbers, postal codes and similar information.

Speech was collected remotely through a dedicated web-based recording platform. Each complete recording is accompanied by an orthographic transcription and recording-level metadata.

The corpus is released under the **Creative Commons Attribution (CC BY)** licence and may be used free of charge for both academic and commercial purposes.

Highlights

- **132,656** complete transcribed recordings
- **221.74 hours** of Polish speech
- **210** speakers: 120 female and 90 male
- known speaker ages ranging from **11 to 69 years**
- **704,259** transcribed word tokens
- **57,640** distinct orthographic transcriptions
- **22.05 kHz**, **16-bit PCM**, mono WAV recordings
- speaker-level and recording-level metadata
- available under the **CC BY** licence

Corpus statistics

Property	**Value**
Complete transcribed recordings	**132,656**
Total duration	**221.74 hours**
Speakers	**210**
Female speakers	**120**
Male speakers	**90**
Known age range	**11-69 years**

Property	Value
Speakers with known age	**207**
Speakers with missing age	**3**
Mean known speaker age	**28.4 years**
Median known speaker age	**25 years**
Transcribed word tokens	**704,259**
Distinct orthographic transcriptions	**57,640**
Mean utterance duration	**6.02 s**
Median utterance duration	**5.75 s**
95th percentile of utterance duration	**10.50 s**
Mean words per utterance	**5.31**
Median words per utterance	**6**
Mean recordings per speaker	**631.7**
Median recordings per speaker	**623**
Recordings per speaker	**153-1,663**
Mean speech duration per speaker	**1.06 hours**
Speech duration per speaker	**0.17-2.39 hours**
Audio format	WAV
Encoding	signed PCM
Sampling rate	22,050 Hz
Bit depth	16 bit
Channels	Mono

The values above were calculated from complete metadata records: entries containing a transcription, a positive word count and a positive recording duration. The supplied file-level SQL dump contains 132,736 rows in total, of which 80 are incomplete metadata entries and are excluded from these corpus statistics.

Speech material

SNUV was designed to support ASR systems that recognise spoken spelling and numbers. The transcriptions show several recurring forms of speech material, including:

- spoken Polish letter names and letter sequences,
- spelled words,
- sequences of individually spoken digits,
- numbers pronounced as whole numerical expressions,
- sequences containing several numerical groups,
- utterances combining spelling-related and numerical material.

Parentheses in the orthographic transcription are used to group components pronounced as a numerical expression, for example:

```
(trzydzieści dwa) (dziewięćdziesiąt pięć) (czterdzieści dwa)
```

A sequence of separately spoken digits may be represented as follows:

```
jeden osiem dziewięć pięć dwa cztery
```

Spelled material is transcribed using the spoken forms of Polish letter names, for example:

```
es pe er zet ą eś ći
```

Speaker distribution

The corpus contains recordings from 210 speakers. Of these, 207 have a known age in the supplied metadata. Their ages range from 11 to 69 years, with a mean of 28.4 and a median of 25 years.

Recording distribution

The average complete utterance lasts 6.02 seconds and contains 5.31 transcribed word tokens. Half of the recordings are no longer than 5.75 seconds, while 95% are no longer than 10.50 seconds.

The contribution per speaker is uneven, as expected in a crowdsourced collection. Individual speakers contributed between 153 and 1,663 complete recordings, corresponding to approximately 0.17-2.39 hours of speech.

Metadata

The supplied metadata is organised into two relational tables represented by SQL dumps.

Recording-level metadata

Each file-level record may contain:

- recording identifier,
- speaker identifier,
- filename,
- relative file path,
- orthographic transcription,
- transcription word count,
- file size,
- recording duration,
- audio encoding,
- number of channels,
- sample size,

- sampling rate.

Speaker-level metadata

Each speaker-level record contains:

- internal speaker identifier,
- identifier used by the recording platform,
- age, where available,
- sex,
- declared recording time,
- actual recording time.

The sex values in the original metadata use the Polish abbreviations `k` and `m`, corresponding to female and male speakers respectively. An age value of zero occurs for three speakers and is treated here as missing information rather than as a literal age.

Recording procedure

The corpus was created using a crowdsourcing approach. VoiceLab was commissioned to develop a web-based application for recording speech and to recruit participants. Contributors recorded remotely and were paid according to the amount of speech supplied, up to approximately 100 PLN per participant. VoiceLab also processed the recordings to provide the required recording quality.

Because contributors used an online collection platform, SNUV represents remotely collected speech rather than recordings made exclusively in a controlled studio environment.

Corpus organisation

The corpus is organised by speaker. The example distribution contains speaker directories with paired files such as:

```
2k36_23lat/
|-- 1_23lat.wav
|-- 1_23lat.txt
|-- 2_23lat.wav
|-- 2_23lat.txt
`-- ...
```

The WAV file contains the speech recording and the correspondingly named UTF-8 text file contains its orthographic transcription. The full distribution also includes SQL metadata dumps describing the speakers and individual recordings.

Technical characteristics

- file format: WAV
- audio encoding: signed PCM
- sampling rate: 22.05 kHz
- sample size: 16 bit
- channels: mono
- transcription format: plain-text orthographic transcription
- metadata format: SQL database dumps

Intended applications

SNUV is intended primarily for:

- automatic speech recognition,
- spoken spelling recognition,
- spoken digit and number recognition,
- voice-driven entry of names and identifiers,
- telephone-number, account-number and code recognition,
- call-centre and customer-service speech technologies,
- evaluation and benchmarking of ASR systems for short utterances.

Availability

The SNUV corpus is distributed free of charge under the **Creative Commons Attribution (CC BY)** licence and may be used for both academic and commercial purposes.

To obtain the database, please contact pelcra@uni.lodz.pl.

Reproducibility

The statistics and figures on this page were generated directly from the supplied file-level and speaker-level SQL metadata dumps. A Python analysis script is available with the documentation materials. It produces:

- a JSON file containing corpus-level statistics,
- a CSV file containing per-speaker statistics,
- speaker-age, utterance-duration and recordings-per-speaker figures in PNG and SVG formats.

Citation

When using SNUV in research or a commercial product, please cite the accompanying corpus documentation and acknowledge the creators of the resource in accordance with the CC BY licence.