Task 1: Punctuation prediction from conversational language

Piotr Pęzik, Adam Wawrzyński, Michał Adamczyk, Sylwia Karasińska, Wojciech Janowski, Agnieszka Mikołajczyk, Filip Żarnecki, Patryk Neubauer and Anna Cichosz

Voicelab, University of Łódź
Motivation

- Speech transcripts generated by ASR systems typically do not contain any punctuation or capitalization.
- Lack of multimodal punctuation prediction datasets in Polish
- In longer stretches of automatically recognized speech, lack of punctuation affects the general clarity of the output.
- Punctuation improves many NLP downstream tasks e.g. text segmentation, indexing and searching, named entity recognition, uppercasing
- The task focuses on predicting punctuation for audio matched with transcripts of natural conversations.
Task

The goal of the present task is to provide a solution for predicting punctuation in the test set collated for this task.

Task includes prediction of the following punctuation marks: fullstop, comma, question mark, exclamation mark, hyphen, ellipsis.
The test set consists of time-aligned ASR dialogue transcriptions from three sources:

- **CBIZ**: a subset of DiaBiz
- **VC**: transcribed video-communicator recordings
- **Spokes**: a subset of the Spokes corpus: casual conversations which were recorded in everyday communicative contexts
Evaluation

Submissions are compared with respect to the weighted average of F1 scores for each punctuation sign.

- Per-document score
- Global score per punctuation sign $p$
Results

Task 1: Punctuation prediction
- Oskar Bujacz (83.30)
- Michał Pogoda (82.33)
- Jakub Pokrywka (71.44)