WIKIMEDIA SVERIGE

The Wikispeech Speech Data Collector

Collect vast amounts of freely licensed speech data through crowdsourcing

WIKISPEECH

KTH VETENSKAP OCH KONST

STTS – Speech Technology Services

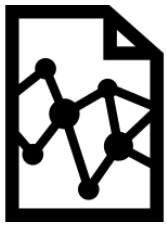
STÖDS OCH FINANSIERAS AV PTS TÄVLING "INNOVATION FÖR ALLA"
pts.se/innovation
PTS

The Wikispeech Speech Data Collector

Collect vast amounts of freely licensed speech data through crowdsourcing

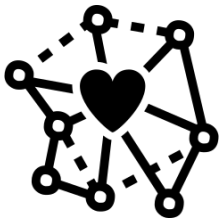
In the project, we will develop tools that will make it **easy** for anyone to contribute.

Depending on the interest and knowledge of the contributor, this can for instance be by recording their own voice or annotating speech audio with linguistic information.



The speech data will be available for anyone who wants to use it, from researchers to product developers to language preservers.

For AI development and research, this data will be of immense value to enhance **Natural-Language Processing** (NLP).



Speech technology applications require speech recordings, often in large amounts and with some linguistic information. Collecting this data is expensive, which is why it is not viable for commercial actors to share.

Since our project will result in a **free and open resource**, we can collect data not only for languages that are the most profitable for commercial products.

Compounded with a close relation to the big, global network of **Wikimedia** volunteers, we will be able to collect data for languages that have little or no resources today.

We will also work towards having a variations of speakers within a given language. This will enable end products derived from this resource will be usable by as many people as possible.

WIKIMEDIA SVERIGE

With support from:

WIKIMEDIA DEUTSCHLAND

Dyslexiförbundet