WIKIMEDIA SVERIGE

The Wikispeech Speech Data Collector

Collect vast amounts of freely licensed speech data through crowdsourcing

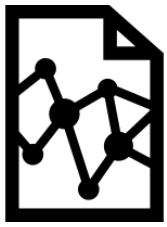
WIKISPEECH

A project by **Wikimedia Sverige**, in cooperation with **KTH Royal Institute of Technology** and **Södermalms talteknologiservice AB**, and with support from **Wikimedia Deutschland**, **Mozilla Foundation** and the **Swedish Dyslexia Association**.

Supported and financed by the the **Swedish Post and Telecom Authority's** competition "Innovation for everyone": http://pts.se/innovation.

The Wikispeech Speech Data Collector

Collect vast amounts of freely licensed speech data through crowdsourcing

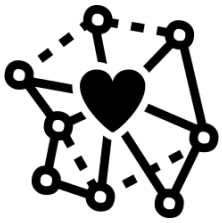


In the project, we will develop tools that will make it **easy** for anyone to contribute. Together we will make **Wikipedia** more accessible for those who cannot read.

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 develop **computers' understanding of language**.



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 where little or no resources are available 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