The Wikispeech Speech Data Collector

A crowdsourcing solution for collecting freely licensed speech data

ANDRÉ COSTA, SEBASTIAN BERLIN (WIKIMEDIA SVERIGE)

WIKIMEDIA SVERIGE

INTRODUCTION

The **Wikispeech** project aims to create an open source text-to-speech tool. Wikispeech will make the information on Wikipedia and its sister projects accessible to a greater audience, such as people with visual impairments, dyslexia, who are illiterate or learn better by listening.

Speech technology applications require speech recordings, often in large amounts and with some linguistic information. Collecting this speech data is expensive, which is why it is often not viable for commercial actors to share. The **Wikispeech Speech Data Collector** will be a free and open toolkit, which will allow collection of data beyond the languages that are the most profitable for commercial products.

The unique multilingual crowdsourcing potential of the global Wikimedia community makes our movement perfectly positioned to tackle this challenge.

KTH VETENSKAP OCH KONST

Wikispeech to Date

Wikispeech started in 2015 as an open source text-to-speech solution for MediaWiki, with focus on Wikipedia. In addition to reading the text out loud it is meant to support community improvements to how the text is read out.

Text-to-speech is currently available for English, Arabic and Swedish. Mechanisms for community improvements are in development. A developer version of the tool is available for testing at wikispeech.wmflabs.org.

STTS

Speech Technology Services

The Speech Data Collector

During the development of Wikispeech we discovered that the availability of free and open speech data was limited, especially beyond the larger European languages.

To help remedy this we will develop tools that facilitate crowdsourced creation of speech data. Along with the tools we will develop methodology and toolkits for volunteer driven crowdsourcing events.

The freely licensed tools will enable the creation of large amounts of freely licensed speech data.

The Value of Freely Licensed Speech Data

Freely licensed speech data is a central component needed for the creation of free speech technology applications such as text-to-speech and speech recognition.

For end users this means that more such applications can be made available, especially for languages where these are scarce or non-existent today. Large amounts of freely available speech data is also valuable for e.g. research, language preservation. Additionally, the speech recordings themselves can be used on other Wikimedia projects such as Wikidata, Wiktionary and beyond. The tools also have the potential to support efforts around oral citations.

For text-to-speech on Wikipedia the speech data, and linguistic knowhow where necessary, will make it possible to add more voices and even new languages. Thereby making the knowledge even more accessible.

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

andre.costa@wikimedia.se