DIGITALISERINGSSTYRELSEN
AGENCY FOR DIGITISATION

GovTech-Program
DENMARK

Challenge statement

This challenge originates from the cooperation between the Agency for Digitisation's and a group of different Danish media's[1], on the utilization of existing media material for the development of Danish speech recognition - a 'Speech-to-Text' Danish language model. The purpose with this challenge is to receive qualified suggestions for the development of language components designed for the use in applied Danish speech recognition – a Danish language model for speech recognition. A Danish language model intended to support the development of AI solutions, with the interest to improve the quality of public and commercial service tasks. This could be automatic voice transcription and dictation used when public servants write journals and process case management, or other use cases such as chat bots, assistance to people with writing difficulties, voice-controlled Internet search and communication with smart devices and speakers, etc.

Speech-to-Text - Danish speech recognition

Under the Danish National Strategy for Artificial Intelligence, the Agency for Digitisation is commissioned to develop and improve access to Danish language resource, aimed to accelerate the development of commercial and non-commercial Danish AI. At the core of the initiative, the ambition is to ensure companies' public access to high quality Danish language resources, and break down barriers for the further development of Danish language technology.

In recent years, the development of speech recognition technology has achieved great progress. In large this relates to advances made in the development of (deep) neural networks, availability of open source software and new methods of gathering large amounts of data for training and testing of algorithms. Unfortunately, this fast-paced development mainly accommodates solutions in English speech recognition and to a less extent languages with smaller populations as Danish.

With this 'Speech-to-Text' challenge, the Agency for Digitisation wish to contribute to the further development of Danish speech recognition. One major barrier is the lack of access to large volumes of high-quality data sets. In cooperation with the group of Danish media's the Agency for Digitisation is working on making large data sets of transcribed and time-coded audio files freely available for companies to use, for the development of AI solutions within Danish speech recognition.

For completing the GovTech program's PoC phase, the Agency for Digitisation in cooperation with the group of Danish media, will provide a sample of 100 hours of transcribed and time-coded audio recordings. With the PoC phase the main interest is to have qualified inputs for the development of language components for a Danish language model and complementary toolkits for the application of Danish speech recognition. Focus is on the creation of value in accommodation of needs from potential user, companies that see a value in these language components by applying them in AI solutions.

To ensure the widest possible user base and application of a free Danish language model and complementary toolkits, it is recommended to use an open-source toolkit such as Kaldi, or look into the potential in HTK[2] from Cambridge University or Sphinx from Carnegie Mellon University. However, these recommendations are not a prerequisite for the completion of a PoC. The most important aspect to the PoC process is to source the best options for development of Danish speech recognition components, which companies can use in the further development of AI solutions that understand Danish.

[1] DR, TV2, JP, Ritzau, Aller, Jysk Fynske Medier, Kristeligt Dagblad, Folketinget, Mediehusene Midtjylland, Børsen, Berlingske, Infomedia og NORDJYSKE Medier.
[2] Hidden Markov Model Toolkit