DIGITALISERINGSSTYRELSEN

GovTech-Program DENMARK

Udfordringsbeskrivelse

Denne udfordring udspringer af Digitaliseringsstyrelsen samarbejde med en gruppe af danske mediehuse[1], om nyttiggørelse af eksisterende mediemateriale til udvikling af dansk talegenkendelse – en 'Tale-til-Tekst' dansk sprogmodel. Formålet med udfordringen er, at få skabt et kvalificeret bud på, hvordan der kan udvikles et sæt sprogkomponenter, som kan styrke talegenkendelse på dansk – en dansk sprogmodel til talegenkendelse. En dansk sprogmodel der kan understøtte en bred udvikling af AI-løsninger med potentiale til at effektivisere og forbedre kvaliteten af offentlige og kommercielle serviceopgaver. Herunder eksempelvis automatisk transskribering og diktat til journaler og sagsbehandling, chat-bots, assistance til personer med skrivevanskeligheder, stemmestyret internetsøgning og kommunikation med smart devices og speakers mv.

Tale-til-Tekst - Dansk talegenkendelse

Under *National strategi for kunstig intelligens* har Digitaliseringsstyrelsen ansvaret for at udvikle en '*Fælles dansk sprogressource'* med et formål om, at forbedre vilkårene for udviklingen af danske sprogteknologiske AI løsninger, både de kommercielle og de non-kommercielle. Initiativet skal lette virksomheders adkomst til gode danske sprogressourcer, og nedbryde barrierer for at virksomheder kommer hurtigt i gang, med at udvikle sprogteknologiske AI løsninger på dansk.

I de senere år er der opnået store fremskridt i udviklingen af services og IT-løsninger, hvor talegenkendelse indgår. Dette er stort set relateret til fremskridt med neurale netværk, og tilgængeligheden af open source-software, der indeholder metodisk afprøvede algoritmer og nye metoder til indsamling af store mængder data til træning og test af algoritmer. Desværre går udviklingen stærkere på engelsk og andre hovedsprog end på dansk.

Digitaliseringsstyrelsen ønsker med Tale-til-Tekst udfordringen at bidrage til udvikling inden for danskproget talegenkendelse. En af de store barrier i den forbindelse er virksomheders adgang til tilstrækkeligt store højkvalitets datasæt. Digitaliseringsstyrelsen samarbejder derfor med de danske mediehuse på, at åbne op for og tilgængeliggøre nogle af de store datasæt, som en række danske medier har, med henblik på at tilbyde AI-virksomheder adgang til en større mængde transskriberet og tidskodet sprogdata, som frit kan anvendes i udviklingen af morgendagens AI løsninger inden for talegenkendelse.

Som en del af udfordringen vil Digitaliseringsstyrelsen i samarbejde med gruppen af danske mediehuse levere et sample af 100 timers transskriberede og tidskodede lydoptagelser til gennemførelsen af GovTech programmets PoC-udviklingsfase. Hvordan udfordringen konkret skal løses er åbent for input fra deltagerne, men det centrale er, at der i PoC-fasen leveres konkrete bud på udvikling af sprogkomponenter til en dansk sprogmodel og toolkits til talegenkendelse. Således skal PoC-processen fokusere på at skabe værdi for potentielle aftagere af sprogkomponenter – de virksomheder som arbejder med at udvikle services og produkter, hvor talegenkendelse indgår.

For at sikre åbne muligheder for at kunne udstille løsninger bredt med åbne snitflader anbefales brugen af et open-source toolkit som fx Kaldi, eller de muligheder der er i HTK[2] fra Cambridge University eller Sphinx fra Carnegie Mellon University. Disse anbefalinger er ingen forudsætning for gennemførelsen af en PoC, det afgørende er en afsøgning af de bedste muligheder for udvikling af talegenkendelseskomponenter, som virksomheder frit kan anvende til at skabe AI-løsninger, som forstår dansk.

[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