INSTITUTIONEN FÖR DATAVETENSKAP | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2026-06-10

EXAMENSARBETE Breaking Language Barriers in Real Time: On-Device AI-Driven Speech Translation
STUDENTER Adrian Steene, Lucas Wittich
HANDLEDARE Pierre Nugues (LTH), Zebastian Karra-Krüger (Axis), Dan Lundgren (Axis)
EXAMINATOR Xuan-Son Vu (LTH)

En värld utan språkbarriärer

POPULÄRVETENSKAPLIG SAMMANFATTNING **Adrian Steene, Lucas Wittich**

Vi vill ta översättningen närmare den plats där samtalet sker. Många tal- och översättningstjänster bygger idag på molntjänster, men det gör systemen beroende av nätverk, externa servrar, och potentiellt känslig datadelning. I vårt arbete testar vi om en hårdvarubegränsad enhet kan ta emot tal, skriva ned det och översätta det helt lokalt.

Språkbarriärer märks mest när människor behöver förstå varandra direkt. Då räcker det inte alltid att översätta menyer eller färdiga texter i förväg. Om kommunikationen sker med tal behöver tekniken kunna följa samtalet tillräckligt snabbt för att stödja en pågående interaktion.

Idag sker mycket taligenkänning och översättning i molnet. Det gör det möjligt att använda stora och kraftfulla AI-modeller, men det innebär också att systemet blir beroende av uppkoppling, externa servrar och överföring av potentiellt känslig data. Vi ville därför undersöka om mer av arbetet kan ske där samtalet faktiskt händer.

Vårt system arbetar i två steg. Först gör taligenkänningen om ljud till text. Sedan översätter maskinöversättningen texten till ett annat språk. Uppdelningen gör idén enkel att förstå, men den ställer höga krav på helheten. Om första steget hör fel får översättningen ett sämre underlag. Om andra steget är för långsamt eller osäkert spelar det mindre roll att taligenkänningen fungerar bra.

Vi testade systemet på sex europeiska språk i en virtuell maskin med begränsad hårdvara. Resultaten visade att lokal taligenkänning kan fungera snabbare än realtid under de testade förutsättningarna, om modellen och körningen optimeras rätt. Den bästa varianten kunde vanligtvis köra långt snabbare än realtid, men hela systemet bromsades fortfarande av översättningen.

Vi testade även LoRA-baserad finjustering. I stället för att ändra i den stora originalmodellen fryser man den helt och lägger till ett litet, effektivt extra lager som tränas. Det kunde förbättra vissa svagare språkpar, men anpassningen behövde balanseras så att andra översättningar inte försämrades.

Vårt arbete visar att lokal AI-driven talöversättning är tekniskt möjlig i vissa fall, men den är ännu inte färdig för praktisk användning. För att komma vidare behövs framför allt snabbare översättning, större minnesmarginaler och fler tester i realistiska miljöer.