Resultatrapport
Stemmestyring i multimodal dialog

Skrevet av: Morten Tollefsen

Sist oppdatert: 17.1.12

Bakgrunn og prosjektmål

I forprosjektet Stemmestyrt interaksjon (STEMINT) spurte vi ulike fokusgrupper om behovet for talegjenkjenning. Som en følge av konkursen i Nordisk Språkteknologi (NST) i 2003 ble satsningen på kommersiell, norsk taleteknologi sterkt redusert. Mange funksjonshemmede hadde store forventninger til NST sitt arbeid, og da firmaet gikk konkurs ønsket mange at noen skulle jobbe videre med spesielt talegjenkjenning. STEMINT ble avsluttet den 15. oktober 2006, og la grunnlaget for en søknad om et hovedprosjekt med oppstart i 2007 - Stemmestyring i MUltimodal DIalog (SMUDI).

Den viktigste hypotesen i SMUDI-prosjektet har vært at multimodale brukergrensesnitt med nødvendig redundans, kan benyttes for å ta høyde for at mennesker har ulike preferanser, forutsetninger og behov. Vi ønsket å finne ut om stemmestyring og auditiv feedback i multimodale dialogsystemer kan muliggjøre nye bruksområder, eller gjøre eksisterende tjenester tilgjengelige for brede brukergrupper. Tale kan eksempelvis brukes som alternativ til mus/tastatur for mennesker som har manglende/redusert håndkontroll, når hendene er opptatte eller når oppgaver skal utføres i mørket. Den overordnede FoU-utfordringen i SMUDI-prosjektet har altså vært å finne ut om multimodale brukergrensesnitt kan brukes for å realisere universell utforming i nye produkter og tjenester. Prosjektets hovedmål ble formulert slik:

> *Utvikle multimodale grensesnitt med nødvendig redundans for å sikre universell tilgjengelighet til styring av klientutstyr og nettbaserte dialogsystemer.*

Multimodalitet har vært vesentlig i SMUDI. Det er likevel viktig å presisere at talegjenkjenning har vært det mest fokuserte satsningsområdet. Vi har imidlertid ikke hatt ambisjoner om at talegjenkjenning nødvendigvis skal erstatte eksisterende interaksjon. Det som har vært viktig i SMUDI, har vært å gi nye brukergrupper muligheter for å styre teknologi, og å gi andre mennesker et supplement til eksisterende hjelpemidler og standardutstyr.

FoU-aktiviteter og resultater

Følgende FoU-aktiviteter/milepæler ble definert i SMUDI:

1. Velge teknologi for norsk talegjenkjenning og talesyntese.
2. Definere funksjonalitet/basisvokabular for styring av PC.
3. Implementere stemmestyring i multimodal interaksjon for PC.
4. Integrere piloter for stemmestyring i multimodal interaksjon med eksisterende hjelpemiddelteknologi.
5. Utrede muligheten for godkjenning av produkter som tekniske hjelpemidler.
6. Definere et rammeverk for nettbaserte multimodale dialogsystemer.
7. Implementere piloter på nettbaserte multimodale dialogsystemer
8. Utrede løsninger for kommersialisering av teknologi og prosjektresultater.

Punktene 1-5 og 8 over ble fokusert rundt arbeidet med talestyring av PC. Det viste seg å være veldig utfordrende å velge teknologi. Lenge hadde vi et samarbeid med IBM. Arbeidet med definering av vokabular etc foregikk delvis parallelt (kartlegging av engelske produkter, utarbeidelse av personas, brukerscenarier etc.). Da det viste seg at IBM ikke hadde egnet teknologi for SMUDI-prosjektet, fant vi en samarbeidspartner i Sverige. Veridict gikk i gang med analyse av data fra NST, og en norsk talemotor ble implementert. Etter dette gikk applikasjonsutviklingen nokså raskt, og vi ga piloten navnet VOMOTE. Til tross for det tekniske starttrøbbelet ble piloten langt mer robust enn det vi strengt talt trengte for brukertestingen i SMUDI, og arbeidet med kommersialisering ble videreført i et prosjekt støttet av Innovasjon Norge (norsk/svensk samarbeid). Produktet ble godkjent som hjelpemiddel av NAV i april 2010. VOMOTE beskrives kort nedenfor.

Det ble jobbet med punktene 6 og 7 tidlig i prosjektet. IBM hadde server-basert teknologi, og vi testet det rammeverket som var disponibelt (telefoni-basert). I forbindelse med VOMOTE hadde Veridict og MediaLT laget Norges første talegjenkjenningsmotor basert på NST-dataene, og også i det nettbaserte dialogsystemet ble denne benyttet. I realiteten hadde vi ikke andre valg, og uttestingen ble derfor gjort med talegjenkjenning på klientutstyret. Gjenkjenningen kan absolutt flyttes til server, noe vi blant annet var med å teste for Nuance med app'er for iPhone (Dragon search og Dragon dictate). Talekommandoer ble altså gjenkjent lokalt, og data hentet fra nettet vha. XML. yr.no ble benyttet og piloten beskrives kort nedenfor.

I tillegg til arbeidet med pilotene er det gjennomført en doktorgrad i SMUDI. Doktorgraden har ikke direkte befattet seg med prosjektarbeidet. Tema for doktorgraden er leksikalsk robusthet i automatisk talegjenkjenning, og doktorgraden vil bli levert og forsvart i løpet av første halvår 2012.

VOMOTE: Med VOMOTE kan du styre en PC bare ved hjelp av stemmen, eller i kombinasjon med andre input-enheter (hodemus, tastatur, mus, brytere, pekerskjerm, ...). Det ble gjort en svært grundig jobb mht. å definere kommandoer, lage personas og brukerscenarioer, og ikke minst gjennomføre både eksperttesting og brukertesting. VOMOTE selges nå som et hjelpemiddel. Tilbakemeldingene har vært varierte, men en ting som går igjen er at veldig mange ønsker seg diktering. VOMOTE kan styre musen, formatere tekst, åpne/lukke programmer, bokstavere tekst og mye mer. Diktering av løpende tekst er imidlertid foreløpig ikke implementert.

Multimodalt brukergrensesnitt (yr.no): Vi ønsket å jobbe med et dynamisk datasett for å gjøre testingen av et multimodalt brukergrensesnitt mer spennende og realistisk. Yr.no tilbyr data i form av XML filer, og værvarslingstjenesten var positiv til vår testing. Vi fikk blant annet omgå enkelte restriksjoner for å gjøre implementering enklere. I yr-applikasjonen kan du be om værvarsling for 10 norske byer med dag og tidspunkt. Du kan også velge ved hjelp av tastatur og/eller mus. Datatekniske hjelpemidler som erstatter mus eller tastatur kan også benyttes, for eksempel brytere. I tillegg kan applikasjonen styres ved hjelp av tale. Resultatene vises på skjerm og kan i tillegg leses opp med syntetisk tale. Leselist kan også benyttes dersom det er en skjermleser på maskinen. Resultatene fra brukertestingen beskrives ikke her, men kort oppsummert kan det konkluderes med at mange foretrekker tale for input, og at de fleste (bortsett fra sterkt synshemmede) vil sjekke resultatet på skjerm.

Prosjektgjennomføring

Prosjektet ble ledet av MediaLT. NTNU har stått for veiledning av doktorgradsarbeidet. Programmering og teknisk arbeid knyttet til data fra NST har blitt utført av Veridict. Øvrige prosjektdeltakere har bidratt i forbindelse med brukertesting, vurderinger knyttet til godkjenning av VOMOTE som hjelpemiddel mm.

Det vil naturligvis alltid være avvik ihht. opprinnelig prosjektplan. I SMUDI prosjektet er alle aktiviteter gjennomført, men til litt andre tider enn opprinnelig skissert. Vi hadde planlagt å teste server-basert talegjenkjenning i et multimodalt grensesnitt, men dette ble gjennomført på en alternativ måte som beskrevet over.

Prosjektet er i all hovedsak gjennomført ihht. oppsatt ressursbruk.

Betydning/nytteverdi

Etableringen av Norsk språkbank skjedde i løpet av SMUDI-prosjektet. Språkbanken er nå lagt til Nasjonalbiblioteket. SMUDI-prosjektet var først ute med å utnytte data fra språkbanken, og den kvalitetskontrollen som ble gjort viser verdien av datamaterialet. Dette er viktig for fagmiljøene, blant annet for å avdekke hva som mangler av norske språktekniske data. For næringslivet demonstrerer den norske talegjenkjenningsmotoren at dataene fra språkbanken er et godt utgangspunkt for videreutvikling. Både for fagmiljøer og næringslivet bekrefter SMUDI-prosjektet at talegjenkjenning er en modalitet som bør vurderes i multimodale brukergrensesnitt for å oppnå universell utforming. Dette er viktig etter hvert som Diskriminerings og tilgjengelighetsloven skal sikre at teknologi skal være tilgjengelig og brukervennlig for alle.

For noen gir talegjenkjenning helt nye muligheter. Vi har for eksempel sett at VOMOTE gir mennesker med lammelser i armene kontroll med bruk av PC etter at dette ikke hadde vært mulig på

mange år. For noen funksjonshemmede vil talegjenkjenning rett og slett kunne føre til større selvstendighet og valgfrihet.

MediaLT selger talegjenkjenningssystemet VOMOTE. Dette er et produkt som primært er utviklet for mennesker med mangelfull eller redusert håndkontroll. Videreutvikling av produktet til eksempelvis å omfatte diktering er kostbart med tanke på den nokså smale målgruppen. Vi vurderer imidlertid fortløpende både teknologiutviklingen og finansieringsmodeller for å få til generell norsk diktering. Dette vet vi er et stort ønske for mange funksjonshemmede, og vi vil derfor prøve å jobbe videre med dette i etterkant av SMUDI-prosjektet.

Konklusjon/oppsummering

Talegjenkjenning er en modalitet som kan bidra til universelt utformet IKT. I SMUDI-prosjektet ble dette demonstrert vha. to applikasjoner: stemmestyring av PC (VOMOTE) og et multimodalt brukergrensesnitt mot værvarslingstjenesten yr.no. Stemmestyring kan gi nye brukergrupper muligheter til å styre teknologi, og for alle vil talegjenkjenning gjøre brukergrensesnittet mer fleksibelt. Ulike modaliteter kan benyttes parallelt eller i forskjellige brukssituasjoner. Det ser ut til at talegjenkjenning kan integreres i brukergrensesnittet uten å øke kompleksiteten. Dette er spennende, siden økt fleksibilitet har en tendens til å gjøre brukergrensesnittet mer komplisert.

I SMUDI-prosjektet har talegjenkjenning vært den mest fokuserte input-metoden. Dette var også utgangspunktet, siden f. eks bruk av mus og tastatur er mer kjent. I tillegg til det store potensiale bruk av menneskelig tale representerer, har vi også avdekket mange utfordringer i prosjektet. Noen av de mest sentrale utfordringene inkluderer: valg av mikrofoner og brytere, når/hvordan talegjenkjenningen skal skrus av og på, støy, dialekter, talefeil og behovet for opplæring.

Økt bruk av talegjenkjenning i multimodale brukergrensesnitt vil gjøre at flere kan benytte de samme IKT-systemene. Dette er viktig, og vi både håper og tror at SMUDI-prosjektet har demonstrert nytteverdien i at mennesker kan styre teknologi med talekommandoer. Til tross for at talegjenkjenning innebærer både teknologiske og menneskelige utfordringer, vil økt satsing på stemmestyring i multimodale brukergrensesnitt garantert føre til innovasjon og økt kunnskap som kan redusere mange av utfordringene som finnes i dag.