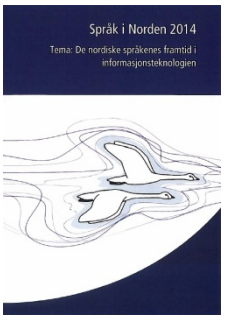
Sprog i Norden

Titel: Norsk språk i eit taleteknologisk perspektiv

Forfatter: Knut Kvale

Kilde: Sprog i Norden, 2014, s. 9-23

URL: http://ojs.statsbiblioteket.dk/index.php/sin/issue/archive



Søgbarhed
Artiklerne i de ældre numre af Sprog i Norden (1970-2004) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Norsk språk i eit taleteknologisk perspektiv

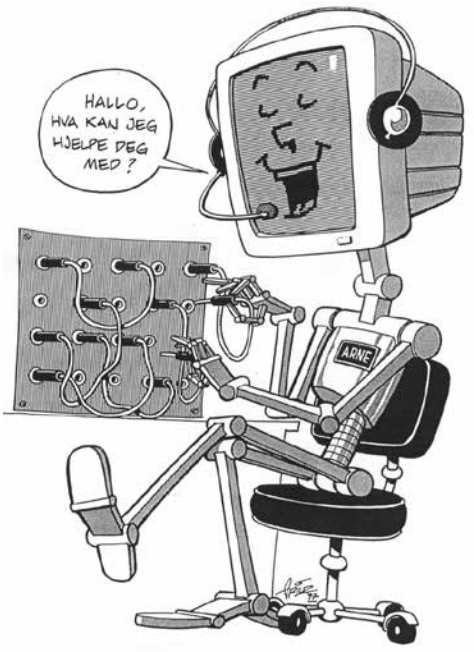
Knut Kvale

Denne artikkelen skildrar fyrst korleis talesyntese og automatisk talegjenkjenning er bygd opp, og drøftar deretter kva for verknader slik taleteknologi kan få for norsk språk. Automatisk talegjenkjenning for generell diktering («tale-til-tekst») er den delen av taleteknologien som kan påverke språket mest. Men programvare for generell norsk diktering finst ikkje enno. Ein av årsakene til dette har vore manglande språkteknologisk infrastruktur i form av ein norsk språkbank, som vart etablert fyrst i 2010. Håpet er at me får samla inn og systematisert språkdata frå heile landet slik at norske talegjenkjennarar kan lærast opp til å kunne tolke dei ulike uttalevariantane i norsk.

Innleiing

Tale er den mest naturlege måten menneska kommuniserer på. Alle er «ekspertar» i å snakke og lytte. Me formidlar tankar og kjensler via eit talt språk utan hjelpemiddel og med liten fysisk innsats. Me kan rusle omkring medan me snakkar, og både augo og hendene er frie til å gjera andre ting. Men når me skal kommunisere med ei datamaskin (PC, nettbrett og andre mobile terminalar), må me halde oss i ro og bruke tastatur og mus eller trykke på skjermen. I mange samanhengar er det difor ynskjeleg å kunne tale til datamaskina (talegjenkjenning) og få talt respons (talesyntese).

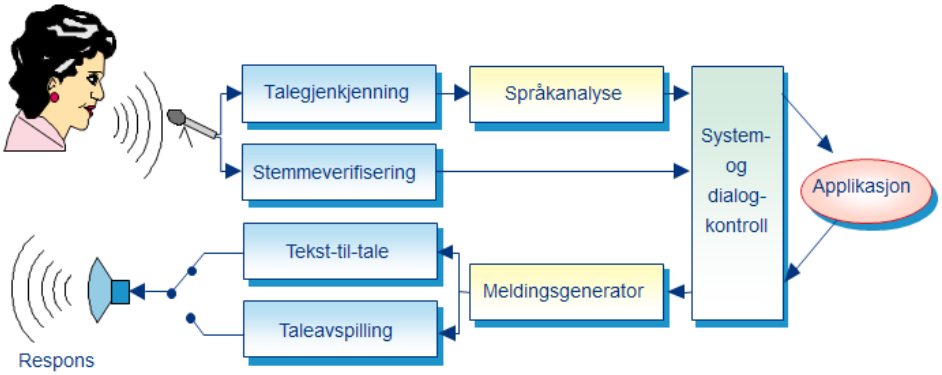
Heilt sidan dei fyrste datamaskinene vart laga, har det vore forska på teknikkar som gjer det mogleg å snakke med datamaskiner, og no har taleteknologien for ein del språk kome så langt at det er mogleg å diktere brev og andre dokument på PC-en, nettbrettet eller mobiltelefonen, eller ein kan spørja om alt frå togruter til aksjekursar. Me har fått *taleføre* datamaskiner som kan hjelpe oss, som illustrert i Figur 1.



Figur 1 Taleteknologi: Ei talefør og hjelpsam datamaskin! (Teikninga sto fyrste gong på trykk i Aftenposten 1. juni 1997, side 30. Teiknaren Arild Midthun har gjeve Språk i Norden løyve til å bruke teikninga).

Kva er taleteknologi?

For å kunne diskutere korleis taleteknologi kan påverke bruken av norsk språk, er det viktig å ha ei felles forståing av kva taleteknologi er. Med taleteknologi meiner me teknikkar for talekoding (talekompresjon), talesyntese, talegjenkjenning og stemmeverifisering. Figur 2 viser typiske komponentar i ei datamaskin som kan snakke og forstå tale. Me seier at datamaskina har eit talebasert brukargrensesnitt.



Figur 2 Talebasert brukargrensesnitt: Datamaskin som forstår tale og som gjev talt respons

Når me snakkar til ei datamaskin med eit talebasert brukargrensesnitt, blir trykkbølgjene i lufta fanga opp av ein mikrofon og klassifiserte som språklydar og ord ved hjelp av automatisk talegjenkjenning. Viss maskina skal kunne gjenkjenne og utføre meir enn ein kommando om gongen, må ein språkanalyse finne grammatisk rette og meiningsfulle setningar frå dei gjenkjende orda. System- og dialog-kontrollen sjekkar om dei gjenkjende orda og setningane gjev eit eintydig oppslag i databasen, og spør eventuelt om att for å sikre rett søk. Responsen frå systemet kan enten spelast av ved å setje saman innlesne frasar (*taleavspeling*), eller det må genererast syntetisk tale direkte frå teksten som blir henta ut or databasen (*tekst-til-tale*).

Figur 2 viser også ein boks med *stemmeverifisering*. Dette er ein teknikk som utnyttar at alle menneske er unike og snakkar forskjellig. Stemmeverifisering kan brukast som ekstra sikkerheit i tilgangskontroll, spesielt ved oppkopling til personlege databasar via telenettet.

Teknikkane for talekoding og stemmeverifisering er ikkje så språkavhengige og vil difor ikkje påverke språkutviklinga. Men det vil kanskje *talesyntese* og *talegjenkjenning* gjere i framtida. Difor ser me nærmare på desse teknikkane nedanfor.

Talesyntese

Talesyntese betyr at datamaskina les opp ein tekst. Me skil mellom (1) talesynteser med eit lite ordforråd og faste meldingar, og (2) talesyntesar som kan lese opp vilkårlege, ukjende tekstar; eit såkalla «tekst-til-tale»-system.

Samanskøyting av faste meldingar

Ein del opplysningstenester nyttar berre nokre få ord og frasar, slik som f.eks.:

> 'Abonnenten har fått nytt nummer', 'Det nye nummeret er', '22' '44' '66' '88'.

For slike tenester kan det løne seg å la ein person lese opp alle orda og dei faste frasane og så lagre dette. Opptaket må merkast slik at den syntetiske talen lett kan genererast ved å kombinere dei lagre orda og frasane. Syntetisk tale basert på innlesne setningar og ord er lett å forstå og høyrest ganske naturleg ut. Men slike system er ikkje fleksible. Viss ein ynskjer å lage ei ny melding eller å utvide ordforrådet, må det gjerast nye opptak med same stemma under same opptaksvilkår. Dette avgrensar mengda informasjon som kan leggjast inn og kor ofte informasjonen kan oppdaterast.

Tekst-til-tale

Ein *tekst-til-tale*-syntese kan lese opp ein vilkårleg tekst. Tekst-til-tale-syntesen er difor svært *fleksibel* og eignar seg spesielt godt for opplesing av ukjent tekst i elektronisk form, slik som elektronisk post, eller for opplesing av tekstlege meldingar som blir endra ofte, slik som nyheiter, vêr- og føremeldingar eller trafikkinformasjon som ligg på nettet. Me kan ringe til slike databasar og få den skriftlege informasjonen lese opp over telefon. Her var Norge tidleg ute: I mai 1998 lanserte Telenor Nextel opplesing av elektronisk post over telefon. Dette var ei av dei fyrste kommersielle tenestene i sitt slag i Europa.

Tekst-til-tale-syntese kan òg nyttast på PC-en, nettbrettet eller mobilen som lesehjelp for blinde og svaksynte eller for folk med ulike lesevanskar. Folk med talevanskar kan bruke tekst-til-tale-syntese som si eiga stemme, som ein "taleprotese".

Det finst mange ulike måtar ein kan syntetisere tale frå tekst. *Figur 3* viser nokre typiske funksjonar i eit tekst-til-tale-system. Teksten må vere i elektronisk form, og kan i prinsippet hentast frå databasar, lesast optisk

med ein skannar, merkast av i eit vindauge på skjermen eller skrivast inn direkte. Systemet må kunne handtere mange typar tekst, slik som forkortingar, tal, datoar, tidspunkt, formlar og tabellar. Dette kallar me *tekstnormalisering*, og det er slettes ikkje ei beintfram oppgåve i norsk.

Eitt eksempel er uttalereglar for tal: Skal talet 1996 lesast som «nitten nitti seks», som «eitt tusen ni hundre og nitti seks», eller som «ein ni ni seks»? Kva med talet "11.10"? Skal det lesast som «elleve komma ti», som «elleve og ein tidel», som klokkeslettet «ti minutt over elleve», eller som datoen «ellevte oktober»? Brøken ¾ skal som oftast lesast som «tre fjerdedelar», men i nokre høve er det ein dato, og då skal det lestast som «tredje i fjerde» eller «tredje april».

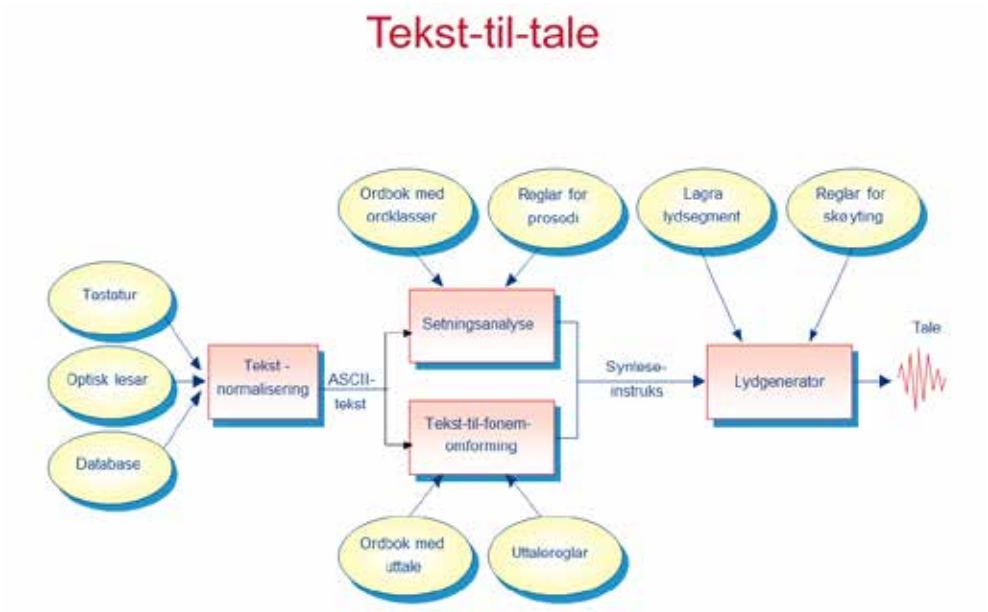
I tekstnormaliseringa kjem ein inn på grunnleggjande spørsmål som å definere kva ei setning er. For inndeling av tekst i setningar, seier me at ei setning sluttar ved punktum etterfølgd av mellomrom og stor bokstav. Regelen ser kanskje rimeleg ut, men vil få problem med setningar som «kvinna var dr. E. Hals». Når det gjeld e-post og nettadresser, må ein bestemme om punktumteiknet skal lesast opp, og korleis det i så fall skal uttalast: som "punktum" eller som "dått"? Talet 11 skal uttalast «elleve», mens «11.» skal uttalast «ellevte». Men kva skal vi gjere når talet står i slutten av ei setning, rett før punktum, som i «Det kom berre 11.»?

Ein må òg lage reglar for korleis forkortingar skal handterast slik at det blir mest mogleg likt vanleg uttale i språket. Her er det fleire alternativ i norsk: Forsvarsalliansen NATO er ei forkorting for «North Atlantic Treaty Organization» og blir uttala som eitt ord, «nato». Forkortinga «NSB», som blir brukt om Noregs Statsbanar, blir uttala som staving: «N-S-B», medan forkortinga «m.a.o.» skal uttalast som «med andre ord».

For å gjere om ein tekst til lydskrift (*tekst-til-fonem-omforming*), slår synteseprogrammet opp i ei ordbok der kvart ord står saman med lydskrifta av ei vanleg uttale av ordet (slik ordet blir uttala isolert i ein valt dialekt). I tillegg trengst uttale av eigennamn, adresser, stadnamn, osb. Viss eit ord ikkje finst i ordboka, må ein ta i bruk generelle uttalereglar for å gjere om teksten til lydskrift. Uttalereglane kan for eksempel vise at bokstavgruppene «skj», «ski», «sj» og «rs» skal uttalast på same måten. I norsk går det an å lage svært lange ord ved å setje saman mindre ord, som for eksempel «gutteinternatskolebestyrar» eller endå verre: «fylkestrafikksikkerhetsutvalgssekretariatslederfunksjonene». Slike samansette ord må systemet prøve å dele opp i mindre ord ved å søkje i ordboka etter grunnforma av dei einskilde orda.

I motsetnad til oss menneske skjønar ikkje datamaskina kva ho sjølv seier (!). Systemet må difor estimere kva for ord det skal leggjast trykk på og korleis setningsmelodien bør vera. I setningsanalysen slår programmet opp i ei ordbok der ordklassa for kvart ord står, slik at for eksempel verb, adjektiv og substantiv blir identifisert. Dette hjelper til med å lage rett uttale og tonelag i setningar av typen «han rota på rommet» og «rota på treet». *Setningsanalysen* blir kombinert med reglar for kva for ordklassar som skal framhevast i setninga og ein modell for setningsmelodi. Viss ein har tilgang til mykje taleopptak som er prosodisk merkt, kan ein utvikle meir avanserte prosodiske modellar og dermed kunne syntetisere meir naturleg tale.

Ut frå *tekst-til-fonem-omforminga* og *setningsanalysen* kjem ein synteseinstruks som inneheld ei rekkje med fonem, samt varigheit, intensitet og grunntone for kvart fonem. Så blir dei *lagra lydsegmenta* henta inn og modifisert med omsyn på synteseinstruksen, før dei blir skøytte saman til lengre lydsekvensar og sende til *lydgeneratoren* – og ut kjem den syntetiske talen.



Figur 3 Blokkskjema av eit tekst-til-tale-system.

Automatisk talegjenkjenning

Med *automatisk talegjenkjenning* meiner me at datamaskina enten skriv det me seier (*"tale-til-tekst"*), eller at ho gjer det me munnleg ber ho om (*"kommandostyring"*). Dette gjer at me kan kommunisere med ei datamaskin utan å bruke tastatur eller mus, noko som er spesielt nyttig for folk med nedsett funksjonsevne i armar og fingrar.

Med automatisk talegjenkjenning treng me ikkje lenger å vere fysisk til stades nær ei datamaskin for å styre ho, fordi me kan gje munnlege kommandoar til datamaskina tvers over rommet eller via telefon. Sidan me kan snakke med datamaskina, blir synet og hendene våre frigjorde til andre formål, noko f.eks. legar utnyttar når dei analyserer mange røntgenbilete og dikterer pasientjournalane samstundes.

Me skal no sjå nærmare på problema i automatisk talegjenkjenning, og korleis desse problema blir løyste. Det aller største problemet for automatisk talegjenkjenning er all variasjonen i talesignala som skal tolkast. Det er mykje lettare å lage god automatisk talegjenkjenning i ei personleg datamaskin der gjenkjennaren kan tilpassast kvar einskild brukar, enn å gjenkjenne kva ulike innringarar seier til ei offentleg telefonteneste.

Talegjenkjenning for PC – talartilpassa gjenkjenning

For nokre språk kan ein no kjøpe dikteringsprogram for PC for under tusen kroner. Slike program er utstyrte med programvare som prøver å lære seg den spesielle måten kvar einskild snakkar på. Dette kallast adapsjon eller talartilpassa gjenkjenning. Programma lèt brukaren diktere og redigere dokument heilt utan tastatur og mus. Ein kan snakke utan pause mellom orda i normalt taletempo, det vil seie mellom 100 og 160 ord i minuttet. I tillegg til diktering kan brukaren redigere teksten ved å seie "uthev ordet", "flytt avsnittet til slutten av dokumentet" osv. Tekstredigering med talte kommandoar gjer at brukarane kan kome til å aktivere verktøy og funksjonar i tekstbehandlaren som dei elles aldri ville ha oppdaga at fanst!

Talegjenkjenning over telefon – talaruavhengig gjenkjenning

Talegjenkjenning i publikumstenester som skal kunne forstå alle folk, blir kalla talaruavhengig gjenkjenning, og er mykje vanskelegare enn talartilpassa gjenkjenning. I slike publikumstenester må systemet prøve å avgrense kva det er lov å snakke om. Dette blir løyst ved at systemet styrer dialogen og stiller innringaren konkrete spørsmål. I teletenester er difor

utforminga av dialogen svært viktig. Kommersielle system er ofte bygd opp rundt ein skjemastruktur der innringaren i praksis fyller ut dei blanke felta i dette skjemaet ved å svare på spørsmål frå systemet. Gjenkjennarane kan i dag skilje ut viktige ord frå høflegsfraser og fyllord, slik at når ein innringar seier "kan eg få æh togopplysning for Dovrebanen, takk", blir orda "togopplysning" og "Dovrebanen" gjenkjent, og rett søk blir utført i databasen.

Variasjon i talesignala: Det store problemet for talegjenkjenning

Sjølv om teknologien for automatisk talegjenkjenning har kome langt, er det lenge til me kan snakke med ei datamaskin på same måten som me snakkar med kvarandre. Det største problemet for ein automatisk talegjenkjennar er all *variasjonen* i talesignala. Kvart menneske er unikt, og alle snakkar forskjellig. Ved nøye analyse av bølgjeformene til lydar ser me at ein person ikkje greier å gjenta nøyaktig same lyden likt to gonger etter kvarandre. Så for datamaskina er kvar lyduttale unik. Dessutan blir realiseringa av språklydane påverka av samanhengen dei blir uttala i. I ei teneste med talebasert grensesnitt for styring av datamaskin over telefon vil det i tillegg vere mange slags bakgrunnsstøy, varierande kvalitet på transmisjonsmedia, ulik kvalitet på mikrofonane, og folk vil variere avstanden til mikrofonane. Den automatiske talegjenkjennaren må difor kunne redusere verknaden av alt som påverkar talesignalet og skilje talesignalet frå støyen.

Teknikkar i talegjenkjenning

Før talegjenkjennaren kan brukast til noko som helst, må han «trenast opp» til å gjenkjenne det vi ynskjer. Vi må bestemme om han skal gjenkjenne setningar, ord eller mindre einingar, ofte kalla «delord». Delord kan vere fonem eller stavingar. Det er svært mange ord i eit språk, men relativt få delord; det er for eksempel omlag 50 fonem i norsk. Talegjenkjenning basert på delord gjer systemet meir fleksibelt. Eit nytt ord kan enkelt leggjast inn i lista av ord som kan gjenkjennast ved å skrive lydskrifta av ordet. Ein slepp altså å trene opp gjenkjennaren på nytt for kvart nytt ord som skal leggjast inn i ordforrådet til gjenkjennaren. Dette kallast gjenkjenning med *fleksibelt vokabular*.

Figur 4 viser eit blokkskjema av ein delordbasert talegjenkjennar. Fyrst blir det digitaliserte talesignalet filtrert og omforma til ein meir kompakt

representasjon. Talesignalet endrar seg relativt seint slik at eit tidsintervall på 10 til 30 millisekund kan representerast med ein eigenskapsvektor med 20 til 50 parametrar.

Ved *mønstergjenkjenninga* blir desse eigenskapsvektorane samanlikna med modellane for dei lagra delorda. Det delordet som liknar mest, blir gjenkjent. Sidan realiseringa av ulike fonem kan likne på kvarandre, bør det takast vare på dei kandidatane som likna nest mest også. På denne måten blir det generert eit nettverk av delord. Det er vanleg å bruke statistiske modellar for delorda. Parametrane i desse modellane blir estimerte («trente») på grunnlag av eit kjent talemateriale («treningssett»).

Ved *leksikalsk omforming* blir nettverket av gjenkjende delord sett saman til ord. Vokabularet til gjenkjennaren står i ei ordliste saman med lydskrift av forventa uttale av kvart ord (slik orda blir uttala isolert i ein vald dialekt). Ei slik ordliste bør kombinerast med reglar for koartikulasjonseffektar over ordgrenser. Ei statistisk ordliste inneheld sannsyn for dei ulike uttalevariantane av orda. Som oftast er det fleire moglege ordkandidatar. Den *syntaktiske analysen* kontrollerer om grammatikken i dei gjenkjende ordsekvensane er rette. Viss to ord har same uttale, men ulik skrivemåte, kan denne analysen hjelpe til med å velje det rette ordet.

Ein *deterministisk grammatikk* er bygd opp av reglar for kva ord som kan følgje etter kvarandre, eller han kan innehalde alle setningane som skal kunne gjenkjennast. Ein *statistisk grammatikk* inneheld sannsynet for at to ord følgjer etter kvarandre (kallast «bigram»), eller at tre ord kjem i rekkefølgje (kallast «trigram»).

Viss datamaskina berre skal utføre enkle kommandoar, er syntaktisk analyse unødvendig.

Ei setning kan godt vere grammatisk rett, men likevel vere meiningslaus. Med setninga «jenta var høy», meiner vi at jenta var lang, men «høy» kan òg tolkast som «tørt gras» eller at jenta var rusa på narkotika. For spesielle oppgåver som f.eks. reservering av billettar eller oppslag i databaser, kan det til ei viss grad modellerast kva som er meiningsfullt (*semantisk analyse*).

Figur 4 Blokkskjema av eit automatisk talegjenkjenningssystem

For at det talebaserte brukargrensesnittet skal vere naturleg å bruke, må orda eller kommandoane som kan gjenkjennast, vere naturlege å bruke i den gjevne samanhengen. Sidan det er mange måtar å uttrykkje same meining på, må talegjenkjennaren ha mange synonym i vokabularet sitt.

Det er ikkje naturleg for oss å seie berre eitt ord eller éin kommando om gongen. Om systemet ber oss om å seie ein kommando, vil mange svare med ei heil setning, for eksempel slik: «Eg vil gjerne ha ... kommando ... takk». Talegjenkjennaren må då kunne overhøyre utanomsnakk og berre gjenkjenne den rette kommandoen. Denne teknikken kallar vi *ordspeiding*.

Viss brukaren ikkje ynskjer å høyre på ei lang utgreiing om moglege menyval, må brukaren kunne avbryte maskina og tale «i munnen» på ho. Systemet må difor kunne lytte og tale samtidig.

Utforming av dialogen er viktigast

Det er vanskeleg å samanlikne ytinga eller nøyaktigheita til ulike system for automatisk talegjenkjenning, fordi dei sjeldan blir testa på det same talematerialet eller på dei same oppgåvene. Kor godt ein automatisk talegjenkjennar verkar er òg avhengig av kva for talemateriale han blir trent på. Det einaste som er sikkert, er at alle system for automatisk talegjenkjenning før eller seinare vil ta feil og misforstå det som er sagt. Det som er viktig for brukarane, er korleis dette systemet handterer gjenkjennarfeil slik at brukarane likevel får utført tenestene dei ber om.

For ei bestemt teneste er det difor viktig å utforme ein god dialog slik at brukaren ikkje gjev opp med ein gong talegjenkjennaren gjer feil. Systemet må kunne spørje om att på ein naturleg måte og få stadfesta eller avkrefta ein hypotese om gjenkjent ord er rett. Viss det likevel blir feil, må det vere enkelt for brukaren å rette opp feilen ved å gå eit trinn tilbake i dialogen. Talegjenkjennaren tek difor vare på ord som liknar på det gjenkjende ordet, slik at systemet kan foreslå eit rimeleg alternativ når det tek feil. For publikumstenester i telenettet bør det også vere enkelt å bli sett over til ein kundebehandlar, for eksempel ved å seie «hjelp» eller «operatør» når som helst i dialogen.

Talebaserte brukargrensesnitt som har hatt suksess for publikumstenester i telenettet, er kjenneteikna ved at brukarane oppfattar tenestene som nyttige, dvs. at tenestene er betre enn liknande tenester eller at det ikkje finst alternativ. Tenestene er brukarvennlege ved at dei:

- *er naturleg å bruke:* Dei erstattar ofte tenester med taste-baserte menyar («for alternativ A, tast 1 ...), og har intuitive namn på kommandoane som kan gjenkjennast.

- *er enkle å bruke:* Dei har innbydande grensesnitt og ein forståeleg dialog.

- *er nøyaktige:* Minst 98 % av orda blir rett gjenkjent.

- *har sanntidsrespons:* Brukaren får raskt svar på spørsmåla sine.

- *har gode dialogar:* Målretta dialogar som reduserer verknaden av gjenkjennarfeil.

Ein god dialog bør styre samtala slik at kunden har få svaralternativ for kvart spørsmål, og kunden blir leia til å svare berre på det han blir spurd om. Dette avgrensar søkjerommet for talegjenkjennaren, og dermed blir risikoen for feil mindre.

Store mengder taleopptak trengst

Me har sett at ein statistisk basert talegjenkjennar må kunne modellere delorda, orda og språket:

- *Delorda;* dvs. modellere karakteristiske akustiske eigenskapar ved delorda og modellere variasjonen i realiseringa av dei.

- *Orda;* dvs. finne sannsynet for ulike uttalevariantar.

- *Språket;* dvs. modellere grammatikken ved å finne sannsynet for at ulike ord følgjer etter kvarandre.

For å kunne estimere parametrane i alle statistiske modellane må det gjerast opptak av mykje tale. Desse talesignala må merkast slik at det er lett å finne fram i talematerialet. Talarane må vere eit representativt utval personar med omsyn til kjønn, alder og dialektbakgrunn. I tillegg bør ein gjere opptak i same omgjevnader som ein har tenkt å bruke gjenkjennaren i. Eksempel: For å få eit realistisk treningsmateriale for ein talegjenkjennar som skal brukast til telefontenester, bør opptaka gjerast over telefonnettet med ulike typar telefonar og i ulike omgjevnader og støymiljø.

Utvikling av norsk taleteknologi har lenge lidd under mangelen av ein norsk språkbank. Store språkteknologiske forskingsprogram som for eksempel «KUNSTI- Kunnskapsutvikling for norsk språkteknologi» (2001–2006), hadde til føresetnad at relevante språkressursar for norsk fanst, eller skulle bli utvikla i form av ein norsk språkbank. Men dette skjedde ikkje, noko som blir oppsummert slik i «St.meld. nr. 35 (2007–2008) Mål og meining - Ein heilskapleg norsk språkpolitikk»:

«Men mangel på relevante språkressursar viste seg å vera eit problem for mange av prosjekta i KUNSTI. Til dels måtte det brukast tid og ressursar på å samla inn og utvikla nødvendig infrastruktur i form av språkressursar før ein kunne starta med sjølve forskingsarbeidet.»

...«Alt i alt vart resultatet i KUNSTI-programmet at ein hadde mindre ressursar å bruka til dei primære forskingsoppgåvene. Dette førte i sin tur at til uheldige justeringar av innretning og ambisjonsnivå i fleire av prosjekta. Samla sett måtte ambisjonane for heile programmet endrast. Bakgrunnen var at ein ved planlegging av programmet hadde lagt til grunn at arbeidet med å byggja opp språkbanken ville koma i gang samstundes med programmet»

Fyrst i 2010 vart Språkbanken etablert, sjå oppdatert informasjon på heimesidene:

http://www.nb.no/Tilbud/Forske/Spraakbanken

Taleteknologi og norsk språk

Domenetap

Det er viktig å vere klar over den trusselen om domenetap som norsk språk er utsett for, mellom anna innanfor taleteknologi i dag. Denne trusselen finn me fyrst og fremst på feltet automatisk talegjenkjenning, som er ressurskrevjande å utvikle.

Følgjande eksempel viser problemet: Eit talegjenkjenningsprogram for engelsk diktering på PC kostar eit par tusen kroner, men å «omsetje» programmet til norsk kjem kan hende på om lag 10–20 millionar kroner. Sidan me framleis ikkje kan tilby ei slik programvare på norsk, er det freistande å diktere e-post eller rapportar på engelsk (med norsk aksent) i staden for å skrive det, spesielt når ein veit at dei fleste norske og skandinaviske mottakarane forstår engelsk.

For å unngå at me i framtida må snakke engelsk til taleføre datamaskiner, er det viktig at Norge satsar meir på utvikling av norsk taleteknologi.

Normering

Vil så norsk talesyntese eller norsk talegjenkjenning ha normerande verknad for norsk uttale? Eit tekst-til-tale-system vil neppe påverke uttalen til lyttaren. Me er vande med at folk snakkar forskjellig, og difor forstår me det maskina seier utan at me sjølve endrar uttale. Men eit talegjenkjenningssystem som berre godtek ein spesiell uttale, for eksempel berre uttalen «sju» og ikkje «syv», kan ha innverknad på uttalen til dei som brukar dette systemet.

Ein skal vere klar over at det er produsentane av talegjenkjenningsutstyr som bestemmer kva for uttale datamaskina skal kunne gjenkjenne. Ved gjenkjenning av norske tal kan for eksempel gjenkjennaren lærast opp til å tolke berre ny teljemåte («tjue-to»), berre gamal teljemåte («to-og-tjue» eller «to-og-tyve»), eller kunne tolke både ny og gamal teljemåte.

På same måten er det produsentane av talesyntese som bestemmer korleis datamaskina faktisk skal uttale orda og setningane. Produsentane kan for eksempel velje kva for dialekt som leggast til grunn for uttalen, dei kan bestemme kor trykket skal leggast i ord som «potet», om syntesen skal uttale tjukk l eller vanleg l i ord som «folk», og kva for r-uttale syntesen skal ha.

Idealet må vere å utvikle norsk taleteknologi som ikkje berre er basert på ein bestemt dialekt eller sosiolekt frå Oslo. Taleteknologien må kunne tilpassast måten folk faktisk snakkar. Prinsippet er at taleteknologien må tilpassast menneska – ikkje omvendt. Difor må me samle inn og systematisere tale frå heile landet, slik at den automatiske talegjenkjennaren kan lære seg både skarre-r og tungespiss-r, ulike trykkleggingar i ord, og uttalar med eller utan tjukk l.

Knut Kvale er seniorforskar ved Telenor Research og professor II ved Universitetet i Oslo, institutt for medier og kommunikasjon.

Summary

The paper describes speech technology and discusses the potential consequences this technology may have for the Norwegian language. Unfortunately, Norway has not yet got an automatic speech recognizer for general dictation. One of the reasons for this is the lack of spoken language resources. Automatic speech recognition (ASR) is based on statistical models, and to be able to estimate the parameters in these models a lot of transcribed speech data is needed. In 2010, "Språkbanken" - a language technology resource collection for Norwegian – was finally established. Hopefully the Språkbank will contribute to the development of ASR-systems that understand the variations in spoken Norwegian.