Bruk av norsk talegjenkjenning for virtuelle besøk innen helsesektoren

Anders Ravndal og Terje Kristensen
Institutt for data- og realfag,
ElHoucine Messaoudi
Institutt for sykepleiefag,
Høgskolen i Bergen
Inndalsveien 28, 5020 Bergen
E-mail: tkr@hib.no, a.ravndal91@gmail.com; El.Houcine.Messaoudi@hib.no

Sammendrag

Artikkelen tar for seg hvordan virtuelle besøk kan nyttes innen telemedisin, og hvordan en kan ha nytte av talegjenkjenning på dette området. En web-applikasjon er utviklet for å utføre videosamtaler der talegjenkjenning brukes til diktering av det som blir sagt. Web-applikasjonen er ment for samtaler mellom pasient og helsepersonell.
For å oppnå talegjenkjenning ble det utviklet både akustiske modeller *(skjulte Markov modeller–HMM*) og språkmodeller (*Trigram)* ved hjelp av åpne tilgjengelige ressurser. Web-applikasjonen ble testet ut av sykepleierstudenter ved Høgskolen i Bergen. Sykepleierstudentene deltok også i en spørreundersøkelse for å kartlegge studentenes meninger om nytten av virtuelle besøk.
Ressursene som ble brukt for å utvikle en talegjenkjenner på norsk var imidlertid ikke tilstrekkelige for bruk av diktering i web-applikasjonen (*Word Error Rate* (WER>100 %)). Studentene var positive til bruk av diktering, men hadde ulike meninger om nytten av virtuelle besøk.

Introduksjon

E-helse dreier seg om løsninger innen helsesektoren der en utnytter informasjonsteknologi. En kjent e-helse tjeneste er e-resepten som er en elektronisk tjeneste for overføring av reseptinformasjon. Telemedisin er en underkategori av e-helse. Bruk av telemedisin [12] kan gi muligheter for helsetjenester selv om den medisinske spesialisten ikke er fysisk tilstede. Telemedisinske løsninger har lenge vært en del av helse-Norge. I 1987 ble en avdeling for telemedisin etablert ved Televerket sin forskningsenhet i Tromsø. Denne avdelingen er i dag kjent som *Nasjonalt senter for samhandling og telemedisin*. I løpet av de snart 30 årene som telemedisin har vært en del av helse-Norge, har utvikling av tjenester innen dette feltet gradvis gått fremover. Utbedringen av mobilnett, båndbredde og video- og lydteknologi har gjort telemedisinske anvendelser mer aktuelle. Telemedisin spiller en sentral rolle ved innføring av Samhandlingsreformen [7], som skal gi bedre samarbeid mellom kommune- og spesialtjenester. Med den fremtidige eldrebølgen i tankene er Norge avhengig av nye og innovative løsninger.
Virtuelle besøk mellom pasient og helsepersonell er globalt en av de mest brukte telemedisinske tjenestene [5]. Bruk av talegjenkjenning i applikasjoner kan åpne for at flere kan ta dem i bruk. Å kombinere en dikteringsløsning med virtuelle besøk vil gjøre

This paper was presented at the NIK 2016 conference. For more information see http://www.nik.no/

det mulig for pasienter med nedsatt hørselsevne å lese det som blir sagt, i tillegg til å høre det. Dette vil også kunne bidra til å effektivisere helse-Norge.

Bakgrunnsinformasjon

En talegjenkjenner har et gitt antall ord som skal gjenkjennes. Dette omtales ofte som vokabularet til talegjenkjenneren. En talegjenkjenner som gjenkjenner kontinuerlig tale med et stort vokabular betegnes ofte som *LVCSR* (*Large Vocabulary Speech Recognition*). LVCSR er et vanskelig problem å løse og brukes i systemer der en ønsker å diktere fritekst. En talegjenkjenner med et mindre vokabular brukes ofte i kommandobaserte system. Et slikt system krever mindre data for å utvikle en god modell. Virtuelle besøk handler om å ta i bruk teknologi for å gi økt grad av tilstedeværelse når det kommuniseres mellom to parter. Graden av tilstedeværelse er avhengig av hvilket utstyr en bruker. Virtuelle besøk brukes vanligvis i telediagnostiske løsninger, men kan også brukes i andre telemedisinske løsninger, som f.eks. å sjekke inn hos en hjemmeværende pasient.
Virtuelle besøk byr på fordeler for å nå pasienter som er langt borte, eller helse-personell i en travel hverdag. Det har også vist seg at bruk av virtuelle besøk kan være gunstig å bruke innen hjemmesykepleien [1], [11]. Av kjente løsninger som eksisterer i Norge i dag er *E-konsultasjon* [10] kanskje den mest brukte. Ved hjelp av denne tjenesten kan en pasient sende tekstmeldinger til sin fastlege, med mulighet for å legge til bilder og dokumenter.

Statistisk talegjenkjenning

Den vanligste modellen for talegjenkjenning baserer seg på statistisk modellering. Disse modellene er basert på skjulte Markov modeller (HMM - Hidden Markov Models) [8], [9]. Også kunstige nevrale nett kan brukes til dette. I statistisk talegjenkjenning ønsker en å finne den mest sannsynlige ordsekvensen som har generert den observerte talen. En talegjenkjenner baserer seg på to ulike modeller. En språkmodell som definerer sannsynligheten for at en gitt sekvens av ord forekommer og en akustisk modell som definerer de statistiske egenskapene til lydhendelser. Den første baserer seg på bruk av N-gram, mens den akustiske modellen baseres på bruk av HMM-er. Statistisk talegjenkjenning defineres slik:

$$S = \arg\max \mathrm{P}(s|X) \tag{1}$$

der S er den mest sannsynlige ordsekvensen, gitt observasjonen X og en tilfeldig ordsekvens s. Ved å anvende *Bayes* sin *regel* kan en skrive formelen ovenfor som [15]:

$$S = \arg\max P(s,X)P(X) = \arg\max P(X|s)P(s) \tag{2}$$

Sannsynligheten $P(X|s)$ er sannsynligheten for observasjon X gitt ordsekvensen. Den vil bli beregnet ved hjelp av den akustiske modellen, mens sannsynligheten for ordsekvensen, $P(s)$, vil bli bestemt av språkmodellen.
En språkmodell definerer sannsynligheten for hvordan en sekvens av ord forekommer. For kommandobaserte talegjenkjennere der en bare trenger å gjenkjenne ett og ett ord, kan en finne sannsynligheten for ordet s_i, $(P(s_i))$, med et vokabular av størrelse K ved å beregne $\frac{1}{K}$ for en uniform fordeling. Ellers kan en definere $P(s_i)$ ved å telle antall

forekomster av s_i i et korpus som modellene trenes på. For en kontinuerlig talegjenkjenner er det mer komplisert å beregne $P(s)$. La s være en sekvens av ord, $s = s_1, s_2, \ldots s_N$. Vi kan da beregne $P(s)$ ved:

$$P(s) = P(s_1, s_2, \ldots, s_N) = P(s_1) \times \prod_{n=2}^{N} P(s_n | s_1 \ldots, s_{n-1}) \quad (3)$$

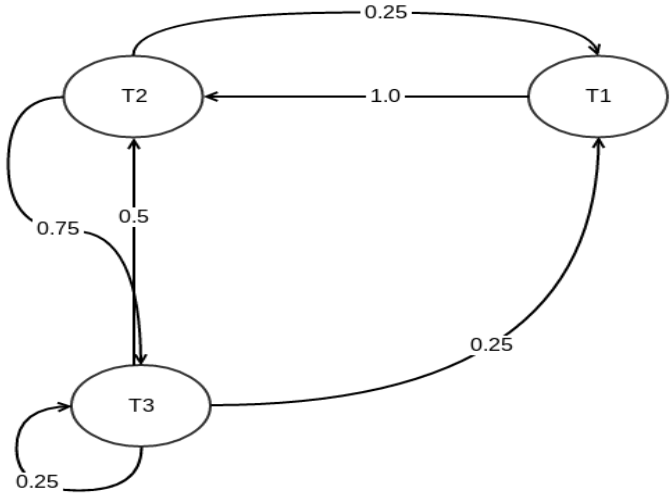
Formelen kan forenkles ved N-gram modellering som blir beskrevet mer utførlig i avsnittet lenger ute.
Den akustiske modellen definerer de statistiske egenskapene til lyd-hendelser. Den gir sannsynligheten for $P(X|s)$ som uttrykkes ved hjelp av HMM-er. I en HMM-basert talegjenkjenning vil en observert sekvens av tale bli knyttet opp mot den HMM-en som passer best til den observerte talen. Dette bestemmes ved å gjøre bruk av Viterbi-algoritmen [8], [9] som omtales lenger ute.

Markov modeller

Markov modeller blir ofte brukt for gjenkjenning av sekvensielle data. Det skilles mellom to typer Markov modeller, Markovkjeder og skjulte Markov modeller. Bruk av sistnevnte viser seg å være nyttig innen talegjenkjenning.

Markovkjede

En Markovkjede er den enkleste Markov modellen en har. Denne modellen er en stokastisk prosess av overganger til forskjellig tilstander. Sekvensen av tilstander er grunnen til navnet Markovkjede. Valget som utgjør den neste tilstanden er bare basert på inneværende tilstand. Dette kalles *Markov egenskapen*. Hver overgang har en viss sannsynlighet, der alle overganger fra en tilstand summeres opp til 1. En tilstand kan også ha en overgang til seg selv.



Figur 1 En Markovkjede

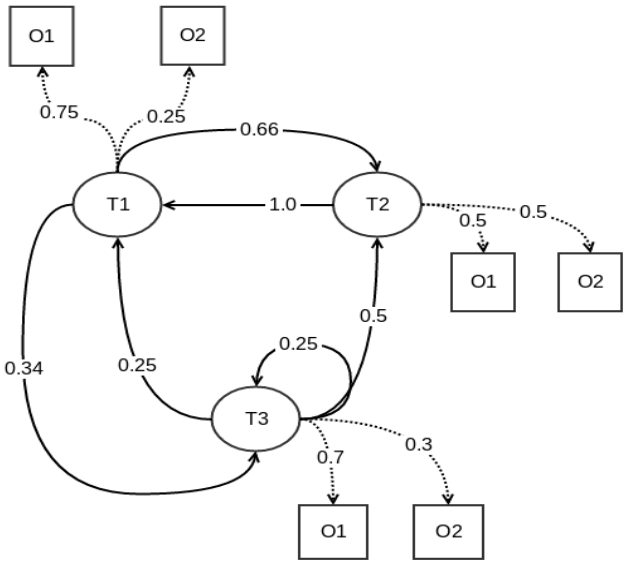
En Markovkjede kan defineres på følgende måte:

- La S være en mengde av N tilstander.
  - N_i representerer den *i*te tilstanden
- La M være en overgangsmatrise av dimensjon $N \times N$
  - Hver radvektor i M representerer overgangssannsynligheten fra en gitt tilstand
  - Summen av elementene i hver radvektor er 1

- La u være en vektor der hvert element er gitt ved sannsynligheten for initialtilstanden
    - Summen av vektorelementene er 1

Skjulte Markov modeller

I en Markovkjede er alle tilstander fullt observerbare, dvs. at man alltid er klar over hvilken tilstand man er i, eller har vært i. En HMM er en utvidelse av en Markovkjede, der hver tilstand avgir et observerbart objekt, mens tilstanden holdes skjult. Det er ingen én til én korrespondanse mellom tilstand og objekt som blir avgitt, derfor kan en ikke være sikker på hvilken tilstand man er i. En HMM består av to sekvenser; en tilstandskjede og en observasjonskjede. I en vanlig Markovkjede vil en gitt sekvens vise en entydig vei i tilstandskjeden. I en skjult Markov modell vil derimot en sekvens av observasjoner kunne ta mange mulige veier gjennom tilstandene.



Figur 2 En skjult Markov modell

En kan definere en HMM på følgende måte:

- La S være en mengde av N tilstander.
- La K være en mengde observerbare symboler k
- La M være en N x N overgangsmatrise som består av overgangssannsynligheter mellom de ulike tilstandene
    - Hver radvektor i M representerer overgangssannsynlighetene fra en tilstand til en annen
    - Et element i matrisen kan representeres som m_{ij}, som er sannsynligheten for en overgang fra tilstand i til j.
    - Summen i hver radvektor er 1, der $0 \leq m_{ij} \leq 1$
- La A være en matrise som sier hvor sannsynlig det er å avgi symbol k i tilstand N_i
    - $a_i(k_j)$ sier hvor stor sannsynlighet der er for å avgi symbol k_j i tilstanden N_i

- La u være en vektor som gir sannsynligheten for opprinnelig tilstand. Da har vi
    - $0 \leq u_i \leq 1$
    - $\sum_{i=1}^{n} u_i = 1$

Markov egenskapen gjelder også for HMM-er. Det vil si at tilstanden ved tid t, $n_1 \dots n_t$, bare er påvirket av tilstanden ved tid $t-1$.

$$P(n_t|n_1^{t-1}) = P(n_t|n_{t-1}) \quad (4)$$

En annen antakelse modellen baseres på er at objektene som avgis er uavhengig av hverandre. Objektet gitt ved tid t, $k_1 \dots k_t$, baserer seg bare på tilstanden man er i ved tid t.

$$P(k_t|k_1^{t-1}, n_1^t) = P(k_t|n_t) \quad (5)$$

Tre problemer

En HMM har tre problemer som lar seg løse. At problemene kan løses på en effektiv måte gjør at HMM-er gunstige for gjenkjenning av sekvensielle data. De tre problemene er følgende:

- Dekoding
    - Gitt en sekvens K med observasjoner, hvilken sekvens, S, (skjulte) tilstander er den mest sannsynlige tilstandskjeden for modellen λ, dvs. finn $P(S, K|\lambda)$
- Evaluering
    - Gitt en sekvens K med observasjoner, hvor stor sannsynlighet er det for at en gitt HMM, λ, vil avgi observasjonssekvensen k, dvs. finn $P(K|\ \lambda)$.
- Trening
    - Gitt en sekvens K med observerbare objekter. Juster modellen λ til en ny modell β som maksimerer sannsynligheten for den gitte sekvensen, dvs. finn $P(S|\beta)$.

Problemene over kan løses ved hjelp av algoritmene Baum Welch, Forward og Viterbi. Om en ønsker en grundigere innføring i disse algoritmene, se [12].

N-gram modellering

En N-gram modell er en sannsynlighetsmodell. Modellen gjør det mulig å beregne sannsynligheten for at en gitt sekvens av N antall symboler vil forekomme. Dette kan en så bruke til å regne ut sannsynligheten for at symbol N vil forekomme gitt sekvensen $N-1$.

Sannsynlighetene bestemmes av at en teller antall forekomster av sekvenser med lengde N over en gitt mengde data. En N-gram modell med en sekvenslengde 1 kalles *unigram*, mens N-gram modeller med lengde på 2 kalles *bigram* og lengde 3 for *trigram*, *osv.*

Sekvens	Unigram	Bigram	Trigram
«...jeg leser en bok…»	jeg, leser, en, bok	jeg leser, leser en, en bok	jeg leser en, leser en bok,

Tabell 1 Eksempler på ulike N-gram for norsk

I avsnittet «Bakgrunnsinformasjon» vil en ikke kunne beregne $P(s)$ i praksis siden antall parametre vokser eksponensielt med antall ord i sekvensen. Vi forenkler derfor uttrykket

ved å innføre den såkalte *Markov egenskapen*: *Fremtidig oppførsel av et dynamisk system avhenger bare av nylig historikk.*

Et N-gram er bare avhengig av N-1 symboler i forkant av symbolet N, dvs. at en N-gram modell er en (N-1)-orden Markov modell. En kte-ordens Markov modell er bare avhengig av de K tidligere tilstandene. Vi omformulerer derfor funksjonen $P(s)$ slik:

$$P(s_1, s_2 \dots N) \approx \prod_{i=1}^{N} P\left(s_i \middle| s_{i-k,} \dots, s_{i-1}\right) = \frac{count(s_{i-k},\dots,s_{i-1},s_i)}{count(s_{i-k},\dots,s_{i-1})} \quad (6)$$

En trenger en mengde sekvensielle data for å trene en N-gram modell. Et N-gram vil være en statistisk representasjon av disse sekvensene. Det er viktig at dataene (sekvensene) man trener N-gram modellen med, er relevant for hvilke problem N-gram modellen skal løse. I en N-gram modell som skal brukes i talegjenkjenning, er det viktig å ha korrekte tekstdata slik at sannsynlighetsfordelingen av sekvensene er tilnærmet lik den som forekommer i virkeligheten. Hvis en skal anvende N-gram modeller på mer spesifikke problemer bør dette også gjenspeiles i treningsdataene.

Utjevning

I virkeligheten forekommer mange forskjellige ordsekvenser. Ofte vil ikke en gitt ordsekvens eksistere i språkmodellen. Dette fordi forekomsten av samme sekvens har vært sjelden eller ikke-eksisterende i treningskorpuset. Språkmodellen vil derfor beregne sannsynligheten til en slik sekvens lik null. Dette er et uønsket scenario. Ved hjelp av utjevning kan en løse problemet ved at språkmodellen også tildeler en sannsynlighet større enn null til slike usette ordsekvenser. Det finnes forskjellige strategier for å løse dette problemet. Slike strategier vil i artikkelen gå inn under begrepet *diskonteringsstrategier*.

Testing

Perplexity (PPL) er den vanligste metoden for å beregne kvaliteten på N-gram modeller. PPL blir estimert ved:

$$PP(z^1 \dots z^n) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(s)}} \quad (7)$$

$$Perplexity(Z) = \sum_{j=1}^{Z} PP\left(z_j^1, \dots, z_j^n\right) \quad (8)$$

der:

- $z_j^1, \dots, z_j^n$ er sekvensen j fra testkorpusfilen
- Z er antall sekvenser i testkorpusfilen.
- $P(s)$ er allerede gitt ovenfor.

Jo mindre verdi PPL i N-gram modellen har, desto større er sannsynligheten for å gjette riktig hva det neste elementet blir (i en språkmodell vil det si å gjette det neste ordet). En kan dermed bruke PPL-verdien til å sammenligne hvor god ulike N-gram modeller er.

Evaluering:

Ordfeilraten (WER-Word Error Rate) [4] er den vanligste metoden en bruker for å evaluere talegjenkjennere. Ved å sammenligne resultatet fra gjenkjenneren med fasit beregnes WER slik:

$$WER = \frac{I+E+S}{E+S+C} \quad (9)$$

der C er antall korrekte ord, I er antall innsetninger som gjøres av gjenkjenneren, E tilsvarer antall endringer en må gjøre for å oppnå korrekt svar, og S tilsvarer hvor mange ord som må slettes for å oppnå et korrekt svar.

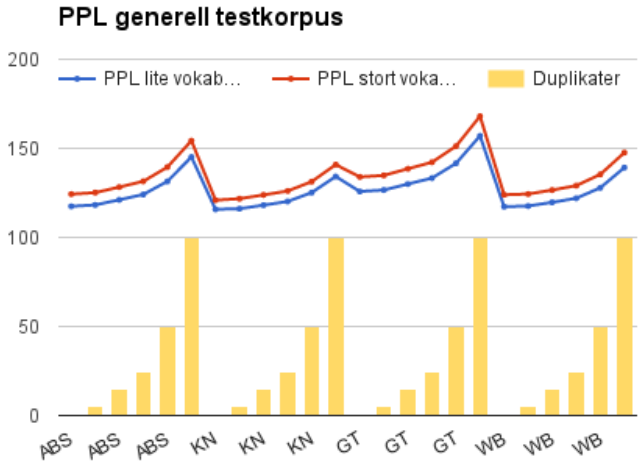
Talegjenkjenning innen Telemedisin

Fra studier innen telemedisin ved *Department of Computer Science* ved *Universitetet i Missouri i USA* [16] har en gjort et forsøk på å anvende teksting av videosamtale under virtuelle besøk ved hjelp av talegjenkjenning. Problemene som avdekkes i denne pilotstudien er at mangel på gode data har stor betydning for hvordan språkmodellen og den akustiske modellen blir dannet. Det er også gjort en studie på tale- og lydgjenkjenningssystemer som skal gjøre hverdagen for eldre tryggere. Det ble installert flere mikrofoner i en hypotetisk leilighet. Tale- og lydgjenkjenningen ble brukt til å oppdage ulike typer lyder som fallende gjenstander, vanlig gange og skrik. Løsningen hadde som mål å oppdage om eldre var i fare, og eventuelt sende et varsel til eksterne enheter.

Eksperimenter og resultater

Resultater fra testing av språkmodeller

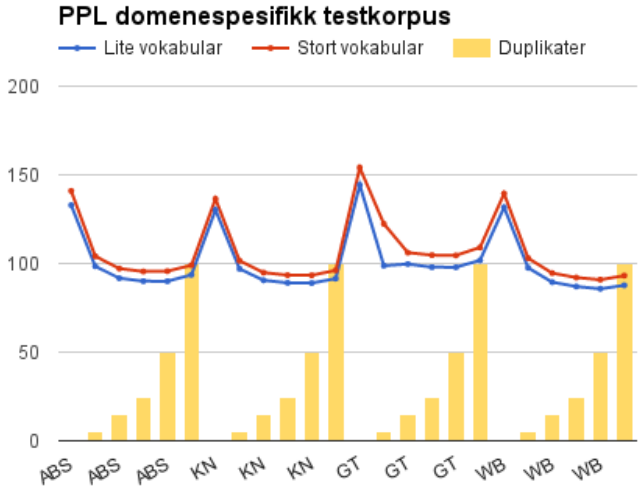
For å utvikle en talegjenkjenner rettet mot det medisinske domenet ble språkmodellene trent på korpuser der linjer som inneholdt medisinske begreper duplisert. I utgangspunktet ønsket en å bruke allerede domenespesifikke korpuser, men det ble ikke funnet fritt tilgjengelige korpuser av slik type. Ved å benytte en dupliseringsmetode ga språkmodellene en større sannsynlighet for å uttale medisinske ord. Språkmodellene ble testet på et generelt testkorpus, og et testkorpus som bare bestod av linjer der en eller flere medisinske ord forekom. Språkmodellene ble også trent på et stort vokabular (ca. 500 000 ord) og et lite vokabular (ca. 60 000 ord). Rammeverket *SRILM* [14] ble tatt i bruk for å trene språkmodellene.



Figur 3: Resultater fra testing av språkmodeller med generelle testdata.

Figur 3 viser PPL- resultatene fra testing på et generelt korpus. Som forventet vokser PPL i takt med antall duplikater i treningsdataene. Under søylene på figur 3 står *ABS*, *KN*, *GT*

og *WB*. Det representerer henholdsvis diskonteringsstrategiene for Absolutt, Kneser-Ney, Good-Turing og Witten-bell.

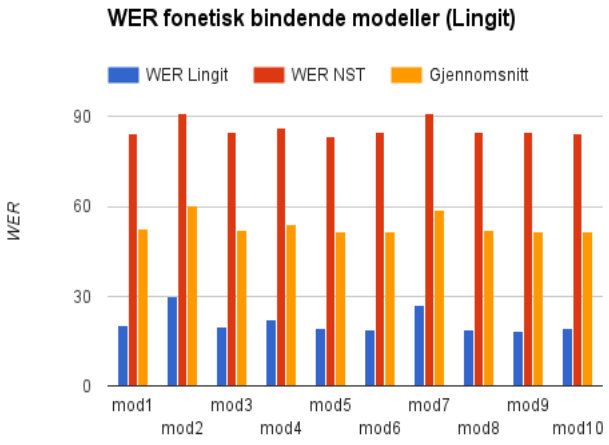


Figur 4: Resultater fra testing av språkmodeller med domenespesifikke testdata.

Figur 4 viser PPL-resultatene fra testing av et domenespesifikk korpus. Her ser en at 50 duplikater er optimalt for samtlige diskonteringsstrategier utenom Absolutt. 25 duplikater gav best resultat. En legger merke til at språkmodellen som er trent på et lite vokabular alltid gir en bedre PPL verdi. Når PPL-verdien beregnes vil ord som ikke er en del av vokabularet til språkmodellen bli ignorert. Med et mindre vokabular blir også sannsynligheten mindre for å gjette galt ord.
Med utgangspunkt i resultatene kan en konkludere med at duplisering av linjer i treningsdataene, som inneholder ord fra et domene, kan ha en positiv effekt ved å skape en mer domenespesifikk språkmodell.

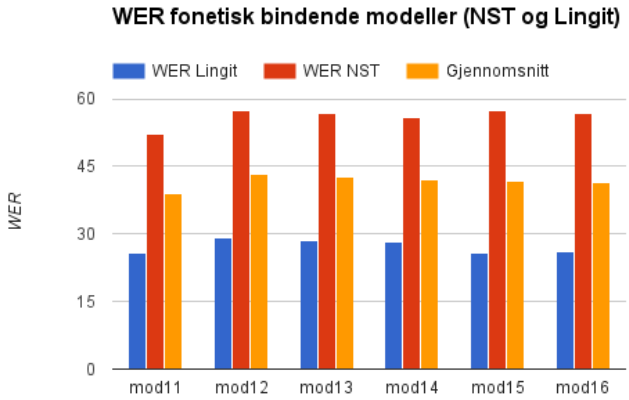
Resultater fra testing av akustiske modeller

Både fonetisk bundne- og kontinuerlige modeller ble utviklet ved hjelp av rammeverket CMUSphinx [3]. I første omgang ble det utviklet fonetisk bundne modeller ved hjelp av *Lingit* [13] sin database. Ved å utføre egentesting av de akustiske modellene, så en at mer taledata var nødvendig for at modellene skal være gode nok for diktering. Taledatabasen til *NST* (NST = Nordisk SpråkTeknologi Holding) [13] ble inkludert i treningen i et forsøk på å skape en bedre akustisk modell. Fonetisk bundne- og kontinuerlige modeller ble så trent med den utvidede databasen.



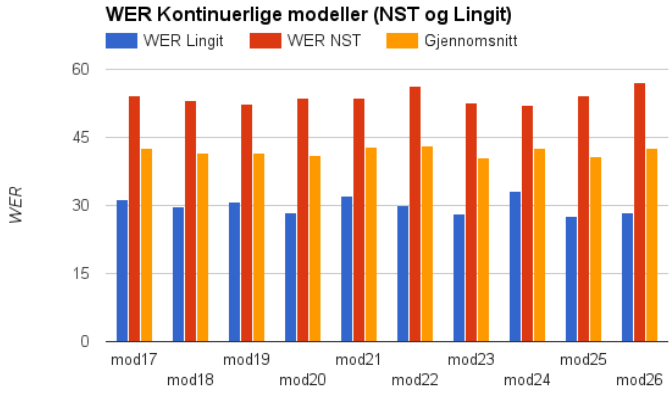
Figur 5 WER-resultater fra fonetisk bundne modeller, trent med Lingit sin taledatabase.

I kontinuerlige modeller har hvert senon et eget sett Gaussiske distribusjoner, og en slik modell kan ha over 100 000 Gaussiske distribusjoner. I fonetisk bundne modeller er det et sett Gaussiske distribusjoner for hvert fonem. En slik modell kan ha ca. 5 000 Gaussiske distribusjoner. Et høyt antall distribusjoner vil gi en økt kompleksitet når Gaussiske blandinger skal beregnes. Fonetisk bundne modeller er derfor en mer effektiv modell enn kontinuerlige modeller. Alle de akustiske modellene ble testet med testdata fra Lingit- og NST sin taledatabase. Det som gjør kontinuerlige- og fonetisk bundne modeller forskjellige er hvordan de Gaussiske blandinger blir bygget opp.
På *Figur 5* ser vi at *mod9* gir best resultater for modellene basert bare på Lingit sine taledata. Denne modellen ga en mindre verdi for WER ved bruk av Lingit sine testdata enn samtlige andre akustiske modeller som ble trent.



Figur 6 WER-resultater fra fonetisk bundne modeller, trent med Lingit og NST sin taledatabase.

På *Figur 6* ser vi at av de fonetisk bundne modellene som ble trent med den utvidede databasen, var det *mod11* modellen som gav best resultater, både for Lingit og NST sin testdata. WER ble beregnet til 25,7% ved testing på Lingit sine data, og 52,3% på NST sin data. Dette ga en gjennomsnittlig WER-verdi på 39%. Modellen ble trent med 6000 senoner og 256 Gaussiske distribusjoner. En kan definere et senon som en enhet for klassifisering av et lydsegment. De 256 forskjellige Gaussiske distribusjonene blir brukt til å bygge forskjellige senoner [12].



Figur 7 WER-resultater fra kontinuerlige modeller, trent med Lingit og NST sin taledatabase.

Ingen av de kontinuerlige modellene som ble trent ga bedre resultater enn *mod11*. Hva årsaken til dette kan være har vi ennå ikke funnet noen forklaring på. Av de kontinuerlige modellene var det *mod23* som ga best resultater med en gjennomsnittlig WER på 40,4%. *mod26* ga best resultater ved testing på Lingit sin testdata med en WER-verdi på 28,3%. *mod19* ga best resultater ved testing på NST sin testdata med en WER-verdi på 52,5%.

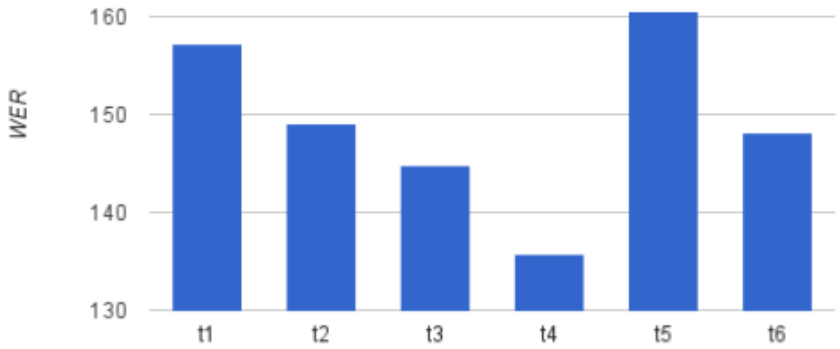
Brukertesting av applikasjon

En test ble utført på et utvalg av sykepleierstudenter med Høgskolen i Bergen. Hver deltaker opprettet en videosamtale med en fiktiv pasient. I tillegg testet deltakerne dikteringsløsningen som web-applikasjonen består av, ved å lese inn 34 setninger. Disse setningene bestod av ingen eller flere medisinske termer som var duplisert i treningsdataen til språkmodellene. I tillegg til den originale testen ble opptakene brukt i senere tester hvor andre språkmodeller ble tatt i bruk.
Testene ble utført med den akustiske modellen *mod11* (se figur 6) og forskjellige språkmodeller. I testene ble dekoderen Pocketsphinx [3] benyttet.

Kjønn	**Alder**	**Hjemsted**	**ID**
Kvinne	21	Voss	t1
Kvinne	21	Tromsø	t2
Kvinne	22	Bergen	t3
Mann	23	Bergen	t4
Kvinne	22	Ålgård	t5
Kvinne	25	Lier	t6

Tabell 2: Informasjon om deltakerne

Ytterlige fire tester viste at en domenespesifikk språkmodell med et lite vokabular ga en økning i gjenkjenning av de medisinske begrepene. Det negative var imidlertid at det også ga en økning i antall feilt gjenkjente medisinske ord. En av hovedårsakene til dette var at ikke alle bøyninger av de medisinske begrepene ble duplisert i treningsdataene til språkmodellene [12].



Figur 8 Forholdet mellom WER og dialekter

På *Figur 8* ser vi at den mannlige deltakeren med bergensdialekt og identitet t_4 gir best resultater. Ålgårddialekten ga dårligst WER-resultat. Grafen viser det gjennomsnittlige WER-resultatene av alle testene som ble gjennomført.

Resultater fra spørreundersøkelse

I tillegg til testing av applikasjon ble det utført en spørreundersøkelse om hvilke meninger sykepleierstudentene hadde om bruk av virtuelle besøk. Undersøkelsen konkluderte med at meningene er delte. På spørsmålet, «tror du virtuelle besøk i fremtiden vil kunne fullt ut erstatte et vanlig legebesøk?», svarte alle sykepleierne at de ikke trodde dette. Begrunnelsen for svaret varierte. En av studentene mente at noen pasientundersøkelser kun er mulig om man er fysisk tilstede. En annen begrunnelse var også at helhetsinntrykket av pasientens tilstand ikke er mulig å oppnå bare gjennom et virtuelt besøk, og at det derfor ikke kan erstatte et virkelig pasientbesøk.

På spørsmålet, «mener du virtuelle besøk vil bedre helse-Norge?», var det delte meninger. To sykepleiere mente at det helt klart ville forbedre helse-Norge. Et av hovedargumentene for dette var at terskelen for å oppsøke helsehjelp ville bli lavere. De mer tvilende studentene mente at en burde undersøke effekten av en slik løsning mer grundig før den ble iverksatt.
På spørsmål om de selv ville brukt en slik tjeneste var også svarene delte. En student mente at det ville passe ypperlig til sin travle hverdag. Legen kunne med god samvittighet gi antibiotika basert bare på studentens sykdomshistorikk og symptomer. Et argument til de som ikke var helt overbeviste, var at de selv ville ha satt pris på den menneskelige kontakten ved et vanlig legebesøk.
På spørsmål om deltakerne selv så for seg å bruke en slik løsning i sin arbeidssituasjon, kom studentene med flere forslag hvor en slik tjeneste kunne tas i bruk. Det ble forslått at en kunne bruke virtuelle besøk til å stille spørsmål rundt behandlingen, og gi pasientene nødvendig forhåndsinformasjon om eventuelle operasjoner de skal gjennomgå. En av studentene som svarte alternativet «tviler» til spørsmålet, mente at det ville være like nyttig å bare ringe pasienter i slike situasjoner i stedet for å ta i bruk videobaserte virtuelle besøk. Når det gjaldt om studentene trodde at deres sykepleierhverdag ville bli mer effektiv ved bruk av virtuelle besøk, svarte fire studenter alternativet «ja, helt klart». Det ble poengtert viktigheten av at tjenesten ikke måtte gå utover kvaliteten. Studenten som stilte seg tvilende til denne tjenesten, mente at mange pasienter ville miste muligheten til en grundig observasjon.
Deltakerne var generelt positive til bruk av diktering ved virtuelle besøk, og at det da ville være enklere for personer med nedsatt hørsel å «fatte» budskapet. For en mer detaljert gjennomgang av resultatene fra spørreundersøkelsen, se [12].

Konklusjon og framtidig arbeid

I artikkelen har vi tatt for oss hvordan en kan nytte talegjenkjenning ved virtuelle besøk. Akustiske modeller og språkmodeller ble utviklet for å gjennomføre det. En Web-applikasjon ble utviklet og testet av sykepleierstudenter ved Høgskolen i Bergen. I tillegg svarte også studentene på en spørreundersøkelse for å kartlegge deres meninger om virtuelle besøk i helsesektoren. Resultatene fra testingen utført av sykepleierstudentene ga en mye høyere WER-verdi enn for testene utført på databasen til NST og Lingit. En av årsakene til dette kan være lydbildet som mikrofonen til studentene skapte.

Vi må erkjenne at vi ikke klarte å skape en talegjenkjenner som var god nok til diktering innen det telemedisinske domenet. NST sin database som ble brukt i treningen av den akustiske modellen, er ennå ikke ferdig utviklet. Vi forsøkte å utføre en opprensking av databasen før trening av de akustiske modellene. Mange «korrupte» filer ble fjernet v.h.a. informasjon fra feilmeldinger som oppstod under trening. Disse filene var som regel lydfiler som bare bestod av støy, og ikke korresponderte med transkripsjonsfilen. NST databasen har et stort potensial. Dersom det hadde blitt utført en grundig opprensing av denne databasen, ville det nok resultert i bedre akustiske modeller. Et treningskorpus som består av samtale mellom pasient og helsepersonell, ville ha bidratt til å utvikle språkmodeller som var bedre rettet mot bruk innen virtuelle besøk. Studentene hadde delte meninger om virtuelle besøk, men så helt klart at det lå et potensial her. Svarene fra spørreundersøkelsen viste også at de var positive til å inkludere diktering i en slik løsning for å nå ut til flere pasienter.

I framtiden må det legges mer arbeid i å skape gode ressurser for å utvikle gode talegjenkjennere på norsk innen det telemedisinske domenet. En kan ta i bruk kunstige nevrale nettverk for å utvikle akustiske modeller. Dette har vist seg å gi en sterk reduksjon i WER [6]. Det må også legges mer vekt på en grundigere brukertesting av applikasjonen, samt en strengere struktur for sikkerhet. Applikasjonen som nå er utviklet er å betegne som konseptuell, og vi har ikke tenkt sikkerhet under utviklingen av den. Mer innovasjon innen talegjenkjenning og større båndbredde vil føre til at teknologi for å realisere virtuelle besøk vil være innen rekkevidde om noen år. Helsenettet [2] gir også et godt fundament for å utvikle en slik tjeneste i praksis.

Referanser

1. **Albertsen, M.H**. Mer enn en telefon - mindre enn et besøk! Masteroppgave, Universitetet i Tromsø, Hansine Hansens veg 18, 9019 Troms, 2012.
2. **Almklov, A**. Stiftelsesdokumen: Norsk Helsenett SF. https://www.nhn.no/om-oss/Documents/styrende-dokumenter/stiftelsesdokument-NHN.pdf , 2009. Accessed: 2016-03-31.
3. **CMUSphinx**: Akustiske modeller. http://cmusphinx.sourceforge.net/. Accessed 2016-03-19.
4. **Evermann, G**. *Minimum word error rate decoding. Cambridge University*, 1999.
5. **Fong, B., Cheuk, A., Fong, M., and Li, C.K.** Telemedicine technologies:Information technologies in medicine and telehealth. John Wiley & Sons, 2011.
6. **Gaida, C., et al.** *Comparing open-source speech recognition toolkits*, 2014.
7. **Helse og Omsorgsdepartementet**. Samhandlingsreformen. Rett behandling - på rett sted til rett tid. *Helse-og omsorgsdepartementet*, Oslo, 2009.
8. **Huang, X., Avero A., Hon, H.** Spoken Language Processing. A Guide to Theory, Algorithm, and System Development. Microsoft Research. Prentice Hall PTR, New Jersey, 2001.
9. **Jurafsky, D., Martin, J.A**. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, New Jersey, 2000.
10. **Kom i kontakt med fastlegen på nett**. https://helsenorge.no/kontakt-fastlegen/kom-i-kontakt Accessed: 2016-03-19.
11. **Lunde Huse, A.M, Storm, M.** Virtual visits in home health care for older adults. The Scientific World Journal, 2014.
12. **Ravndal, A**. Norsk talegjenkjenning i telemedisin. Masteroppgave i programutvikling, Institutt for data- og realfag, Høgskolen i Bergen og Institutt for informatikk, Universitetet i Bergen, Norge. 2016.
13. **Språkbanken**:http://www.nb.no/sprakbanken/repositorium#ticketsfrom?lang=nb&collection=sbr. Accessed: 2016-04-22.
14. **Stolcke, A.** SRILM - an extensible language modeling toolkit. *Interspeech* , September, 2002.
15. **Young, S., et al.** The HTK book, volume 3.2. Entropic Cambridge Research Laboratory Cambridge, 1997.
16. **Zhao, Y., et o.** An automatic captioning system for telemedicine. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing conference, ICASSP-2006, volume 1. pp 1, 2006.