Nasjonalbiblioteket

Status for norsk talegjenkjenning

Per Erik Solberg, Marie Røsok, Ingerid Løyning Dale, Arne Martinus Lindstad

Sammendrag

Språkbanken ved Nasjonalbiblioteket har testa hvor gode talegjenkjenningssystemer er til å transkribere norsk tale til bokmål og nynorsk, og denne rapporten rapporterer resultatene fra testinga. Vi har utvikla et testsett med taledata fra TV- og radiosendinger fra NRK som vi har transkribert manuelt til bokmål og nynorsk. Ved hjelp av dette testsettet har vi målt kvaliteten på ulike talegjenkjenningssystemer. Ny teknologi og økt tilgang på norske treningsdata for talegjenkjenning har gjort norsk talegjenkjenning mye bedre de siste åra, og testingen viser at de beste systemene transkriberer godt på bokmål og nynorsk. Nasjonalbibliotekets Whisper-modeller, NB Whisper, viser seg å produsere de beste transkripsjonene på bokmål og nynorsk.

Rapporten peker også på områder der det er rom for forbedringer:

1. **Bokmål og nynorsk**: Systemene som støtter begge skriftspråk, produserer av og til nynorskord i bokmålstranskripsjoner eller bokmålsord i nynorsktranskripsjoner, særlig når taleren har en utprega nynorsknær eller bokmålsnær dialekt.
2. **Overlappende tale:** Lydfiler med overlappende tale blir gjennomgående dårligere transkribert enn lydfiler med bare én taler. Overlappende tale er vanskelig for maskiner som for mennesker, men dersom talegjenkjenningssystemer skal kunne transkribere dialoger på en god måte, er det viktig at de klarer å transkribere overlappende tale.
3. **Dialekter:** De beste systemene håndterer ulike dialekter godt, men vi har likevel observert at systemene er bedre på noen dialekter enn andre. Dette gjelder særlig for nynorsktranskripsjon, der nynorsknære dialekter blir transkribert bedre enn bokmålsnære.
4. **Ordrett og ikke-ordrett transkripsjon:** Systemer basert på Whisper-teknologi kan produsere transkripsjoner som ikke er ordrette: De kan gjengi meningsinnholdet uten nødvendigvis å bruke like mange ord som i lydopptaket, og de kan også omformulere. Slike ikke-ordrette transkripsjoner er nyttige for mange formål, for eksempel for undertekst og møtereferater. Whisper-systemene vi har testa, produserer ikke-ordrette transkripsjoner av høy kvalitet. Noen ganger fjerner imidlertid systemene for mye av det som opprinnelig ble sagt.
5. **Hallusinering og fremmede språk:** Vi ser at en noen systemer hallusinerer, det vil si at de produserer transkripsjoner som ikke samsvarer med lydfila overhodet, eller skriver på andre språk enn norsk.

Innholdsfortegnelse

1. Om rapporten og hovedmål

I 2010 ble Språkbanken oppretta ved Nasjonalbiblioteket for å sikre at det fantes gode grunnlagsressurser for norsk språkteknologi og gjennom det sikre at norsk ikke ble hengende etter engelsk i teknologisk sammenheng. I 2019 ble det en fornya satsing på språkbankoppdraget i samarbeid med Språkrådet. Taleteknologi, og ikke minst talegjenkjenning har vært et viktig satsningsområde for Nasjonalbiblioteket siden 2019. Språkbanken har utvikla flere datasett og andre grunnlagsressurser for talegjenkjenning, og Språkbanken deltar i det NFR-finansierte infrastrukturprosjektet SCRIBE om talegjenkjenning av dialoger og spontan tale, det vil si tale som ikke er lest opp fra et manuskript. I tillegg har Nasjonalbiblioteket utvikla flere åpne talegjenkjenningssystem. Dette arbeidet har sammenfalt med et teknologisk paradigmeskifte innenfor talegjenkjenning, som har gjort at talegjenkjenning av spontan tale har blitt betydelig bedre.

I lys av arbeidet med talegjenkjenning i Språkbanken ba Språkrådet Språkbanken i 2022 om å utvikle et testsett for å teste hvor godt talegjenkjenningssystemer håndterer dialekter og spontan tale. Dette testsettet skulle brukes til å finne ut hvor gode systemene er per i dag, og det skulle kunne gjenbrukes seinere for å se i hvilken grad Språkbankens arbeid har målbar effekt. Språkbanken ønska også et slikt testsett for å få en oversikt over hva som fungerer godt i dag og hvor det er forbedringspotensial. Språkbanken utvikla et testsett med 10 timer med manuelt transkribert lyd fra ulike NRK-program. Programmene er valgt ut for å få god variasjon av dialekt, kjønn, programtyper og opptaksforhold. Vi har så splitta lydfilene i testsettet slik at det er én lydfil per setning. Disse setningsdelte lydfilene har vi kjørt gjennom ulike talegjenkjenningssystem og målt kvaliteten på de automatisk genererte transkripsjonene opp mot den manuelle transkripsjonen. Denne rapporten beskriver testsettet og resultatet av testinga.

Hovedmålene med testinga er:

1. Evaluere status for norsk talegjenkjenning
2. Finne ut hva slags feil norske talegjenkjenningssystemer gjør og hva Språkbanken skal bruke tid og ressurser på videre
3. Vise forbedringene i norsk talegjenkjenning over tid

2. Litt om talegjenkjenning

Talegjenkjenning er automatisk transkripsjon av taleopptak, enten i sanntid eller fra innspilte lydfiler. Talegjenkjenningssystemer er helt og fullt maskinlæringsbaserte: Når maskinen konverterer tale til skreven tekst, bygger ikke det på regler skrevet av mennesker. I stedet lærer datamaskinen å gjøre denne konvensjonen selv gjennom å observere mønster i data den blir matet med. Resultatet av denne læringsprosessen er en *maskinlæringsmodell*. Dataene modellen lærer fra, kalles *treningsdata*. Innenfor maskinlæring skiller man mellom *veileda læring* og *uveileda læring*. Ved veileda læring er treningsdataene merka opp med mønster som maskinlæringsmodellen skal lære seg å gjenkjenne.
Talegjenkjenningssystemer bruker for det meste veileda læring: Treningsdataene for

talegjenkjenningsmodeller er lydfiler, og oppmerkinga er ortografiske transkripsjoner av disse lydfilene. Ved uveileda læring har treningsdataene ingen oppmerking. Som vi kommer tilbake til i neste kapittel, fins det talegjenkjenningssystemer som delvis er trent på lydfiler uten transkripsjon eller noen annen form for oppmerking. En utfordring for veileda maskinlæring er at det gjerne fins begrensa mengder med oppmerka data, fordi slik oppmerking som regel krever menneskelig innsats. Dette er også tilfelle for talegjenkjenning: Treningsdataene må gjerne transkriberes manuelt, og det er kostbart å få folk til å transkribere taledata.

Fram til nylig trengte de fleste talegjenkjenningssystemer et uttaleleksikon, en liste med ortografiske ordformer og fonetiske transkripsjoner av ord. Flere moderne talegjenkjenningssystemer klarer seg imidlertid uten en slik ordliste.

3. Modellarkitekturer og data

For at talegjenkjenningssystemer skal fungere godt for ulike stemmer, dialekter, opptaksforhold og formål, må de trenes på transkriberte lydfiler som inneholder slik variasjon. Et talegjenkjenningssystem som hovedsakelig er trent på talere fra østlandet, vil for eksempel ikke fungere godt på talere fra vestlandet. En relativt enkel måte å produsere datasett for talegjenkjenning er å hente tekst fra for eksempel aviser og så be informanter om å lese opp teksten. Talegjenkjenningssystemer som kun er trent på slik opplest tale vil fungere dårlig på tale som ikke er opplest, kalt spontan tale, fordi vi snakker ganske annerledes når vi leser opp en tekst fra når vi snakker fritt. Dessuten er det gjerne mindre dialektvariasjon i opplest tale enn i spontan tale. Det er derfor en stor fordel at talegjenkjenningssystemer som skal brukes på spontan tale, også er trent på slik tale.

I dette kapittelet vil vi først beskrive de viktigste datasettene som fins på norsk. De siste årene har det blitt lansert nye typer maskinlæringsmodeller, nye *modellarkitekturer*, for talegjenkjenning, som fungerer langt bedre enn tidligere systemer. En grunn til at de er bedre, er at de utnytter data som tidligere ikke kunne brukes til å trene talegjenkjenningssystemer, som dermed gir tilgang til mer og mer varierte data. Etter å ha presentert de ulike datasettene for norsk, vil vi kort presentere tre slike modellarkitekturer. Til slutt i kapittelet vil vi beskrive de modellene vi har valgt å teste i dette prosjektet og noen egenskaper ved disse systemene.

Datasett for norsk taleteknologi

NST-datasettet

Selskapet Nordisk språkteknologi (NST) laga på tidlig 2000-tall et stort datasett for norsk talegjenkjenning. NST gikk seinere konkurs, og Språkbanken overtok datasettet. Datasettet har blitt delt i Språkbanken siden 2011. NST-datasettet består av over 500 timer med tale fra nær 1000 talere. Talerne leser opp setninger fra et manuskript. De fleste setningene er fra avistekst.

NST-datasettet var fram til 2021 det eneste åpent tilgjengelige datasettet av en viss størrelse for trening av talegjenkjenning og har trolig blitt brukt i flere talegjenkjenningssystemer. Det er stort nok til å være nyttig som treningsmateriale for et talegjenkjenningssystem. Imidlertid inneholder det kun opplest tale, og er dermed ikke så velegna for utvikling av talegjenkjenningssystemer som skal forstå spontan tale. Transkripsjonene er hovedsakelig på bokmål, men det er også en liten andel som er på nynorsk.

Stortingskorpuset (NPSC)

I perioden 2019-2021 utvikla Språkbanken Stortingskorpuset (kalt *NPSC* fra dets engelske navn, *Norwegian Parliamentary Speech Corpus*). NPSC består av 126 timer med tale fra Stortinget og transkripsjoner skrevet av ansatte i Språkbanken (Solberg & Ortiz, 2022). Stortingsdata egner seg godt til datasett for talegjenkjenning av flere grunner: Dataene er åpent tilgjengelige, det fins gode metadata om talerne, det er en god blanding av opplest og spontan tale på Stortinget, og det er god dialektvariasjon. Ca. 13% av transkripsjonene er på nynorsk. I prosjektet valgte man å transkribere fra grunnen av etter detaljerte retningslinjer istedenfor å bruke teksten i de offisielle stortingsforhandlingene. Stortingsforhandlingene er nemlig ikke ord-for-ord-gjengivelser av det som blir sagt i stortingssalen, og de mest brukte talegjenkjenningsarkitekturene da NPSC ble laga, trengte ordrette data.

Det utvida stortingskorpuset (SSC)

Som vi skal se, har det skjedd et teknologisk skifte de siste åra, som gjør det mulig å trene talegjenkjenningssystemer på taledata som ikke er transkribert ordrett. På grunn av dette valgte Språkbanken i 2022-2023 å lage et utvida stortingskorpus (kalt *SSC* etter det engelske navnet *Stortinget Speech Corpus*) med transkripsjoner henta fra de offisielle stortingsforhandlingene (Solberg et al., 2023a). Språkbanken brukte en metode som tidligere var brukt på kroatiske parlamentsdata (Ljubešić et al., 2022) for å ekstrahere transkripsjonene. SSC er på over 5000 timer, som er flere ganger mer enn alle de øvrige, åpne datasettene for talegjenkjenning til sammen. Transkripsjonene er ikke alltid ordrette.

NB samtale

NST, NPSC og SSC mangler samtaler. For å sikre at det også fins åpne taledata med samtaler, utvikla Språkbanken i 2023 NB samtale, som består av transkriberte podkaster og live-arrangementer. Transkripsjonene er laga av ansatte i Språkbanken.

Norsk talestyringskorpus (NVCC)

Norsk talestyringskorpus (*Norwegian Voice Control Corpus*; *NVCC*) er utvikla av Språkbanken i 2020-2022 som treningsdata for prateroboter. Det består av oppleste ytringer beregna på slike roboter. Det har også en liten del med ytringer som ikke er opplest. Dataene er lest av 11 personer med ulik dialektbakgrunn.

Datasett	Størrelse	Antall talere	Tale	Transkripsjon	Språk	Hjemmeside
NST	540 timer	~1000	opplest	ordrett	bokmål	lenke
NPSC	126 timer	267	opplest og spontan	ordrett	bokmål og nynorsk	lenke
SSC	5189 timer	729	opplest og spontan	ikke-ordrett	bokmål og nynorsk	lenke
NB samtale	24 timer	69	spontan	ordrett	bokmål og nynorsk	lenke
NVCC	10 timer	11	opplest og spontan	ordrett	bokmål og nynorsk	lenke

Tabell 1: Oversikt over datasett for norsk talegjenkjenning

Viktige modellarkitekturer for talegjenkjenning

Klassiske talegjenkjenningssystemer

Fram til ca. 2020 var de fleste talegjenkjenningssystemene *modulære*: De besto av flere moduler som løste ulike oppgaver. En av modulene i slike modulære systemer er en akustisk modell, en maskinlæringsmodell som predikerer sannsynlige sekvenser av språklyder i en lydfil. Den akustiske modellen er trent på transkriberte taledata. For den akustiske modellen i slike systemer er det viktig at transkripsjonene i treningsdataene er så nøyaktige gjengivelser som mulig av det som blir sagt, inkludert kremting, nølelyder, avbrutte ord o.l. Det er få andre bruksområder for slike helt ordrette transkripsjoner, og derfor kunne man stort sett kun bruke datasett som var utvikla spesielt for trening av akustiske modeller. Da Språkbanken utvikla NPSC, kunne de for eksempel ikke bruke stortingsforhandlingene som transkripsjon, men måtte ansette transkribører og utvikle spesifikke transkripsjonsretningslinjer.

I tillegg til den akustiske modellen har slike systemer et uttaleleksikon, som forteller hvilke sekvenser av språklyder som svarer til ord i språket, og en språkmodell, som sier hvor sannsynlig det er at en transkripsjon skapt av systemet er en setning på språket modellen er laga for. Språkmodellen er trent på tekstdata.

For å utvikle klassiske talegjenkjenningssystemer trenger man altså store mengder ordrette treningsdata, slik som NST-datasettet og NPSC, i tillegg til uttaleleksika og tekstdata for å trene en språkmodell. Med unntak av tekstdataene er dette data som må utvikles spesielt for formålet. Å utvikle slike data med den nødvendige variasjonen er dyrt og krevende, og

tilgangen på data er derfor en flaskehals, spesielt for små språk som norsk. På grunn av dette forsøker en del nyere modellarkitekturer å gjøre seg mindre avhengig av slike spesiallaga ressurser. I tillegg utnytter de bedre moderne, kraftig maskinvare.

wav2vec2

Wav2vec 2.0 er en maskinlæringsarkitektur for talegjenkjenning og andre former for taleprosessering som ble introdusert i 2020 av forskere hos Meta (Baevski et al. 2020). Wav2vec2 bruker *transformers*-teknologien, som har revolusjonert AI-feltet de siste åra, og som også ligger til grunn for systemer som ChatGPT og BERT. Denne teknologien kjører effektivt på moderne maskinvare og er godt egna til å gjenkjenne mønster i språk. Wav2vec2 utnytter fordelene denne teknologien gir, til å minske behovet for kostbare, håndlaga ressurser for talegjenkjenning:

1. Treningen av en wav2vec2-modell foregår i to steg. Først trenes en *forhåndstrent modell* på store mengder taledata uten transkripsjoner. Dette er *uveileda læring*, fordi dataene ikke har transkripsjoner eller noen annen form for oppmerking. Den forhåndstrente modellen tilegner seg viktig informasjon om tale og nyanser i tale fra å ha observert store mengder taledata, men den kan ikke alene transkribere lydfiler. Derfor må den *fintunes,* eller finjusteres, på et transkribert datasett for å lære seg å transkribere. Siden den forhåndstrente modellen allerede inneholder mye informasjon om tale, trengs det mindre transkriberte data for å fintune den enn det som er nødvendig i klassiske talegjenkjenningssystemer. De fleste talegjenkjenningsmodeller basert på wav2vec2 er fintuna versjoner av allerede eksisterende forhåndstrente modeller, så man kan lage gode systemer uten å utvikle en forhåndstrent modell selv.
2. Wav2vec2-arkitekturen trenger ikke noe uttaleleksikon eller noen annen form for ordliste. Den fintuna modellen forsøker å konvertere lydfilen direkte til sekvenser av bokstaver og mellomrom. Systemet trenger heller ikke en språkmodell, men en språkmodell kan med fordel brukes for å gjøre systemet bedre på å stave korrekt.

Wav2vec2 omgår altså behovet for uttaleleksikon fullstendig, og en språkmodell er heller ikke strengt nødvendig, selv om det ofte er fordelaktig å bruke en språkmodell. Systemet trenger også mindre transkriberte taledata for å gi gode prediksjoner enn klassiske systemer. Taledataene som trengs, må imidlertid være tilnærma ord-til-ord.[1] Transkripsjonene systemet produserer, er også ord-til-ord.

En konsekvens av at systemet ikke har noe leksikon eller ordliste, er at det kan stave feil. Dette skjer særlig med ord som det ikke har sett før og som har en uregelmessig ortografi, sånn som fremmedord. Det kan også skje dersom taleren snakker en dialekt systemet i liten grad har sett i treningen. Det blir gjerne færre slike feilstavinger dersom man bruker en språkmodell.

[1] Det fins vellykka eksperimenter der man har trent wav2vec2-modeller på taledata som ikke er helt ordrett transkribert, jf. Ljubešić et al. (2022).

Whisper

I september 2022 lanserte OpenAI, firmaet bak ChatGPT, en flerspråklig talegjenkjenningsmodell som heter Whisper (Radford et al., 2023). I likhet med wav2vec2 bruker Whisper transformers-teknologi, men arkitekturen er ellers ganske forskjellig. Alle treningsdataene til Whisper er transkribert. Med andre ord, systemet bruker kun veileda læring. Imidlertid trenger ikke Whisper treningsdata som er transkribert ord for ord. Den flerspråklige modellen til OpenAI er ikke trent på datasett som er spesiallaga for talegjenkjenning. I stedet har de lasta ned videoer med undertekst fra internett og har trent modellen på det. Siden det fins mye mer teksta videoer på internett enn det fins datasett for talegjenkjenning, er OpenAIs modell trent på mer data enn noe annet talegjenkjenningssystem: 680 000 timer med flerspråklig, transkribert tale. OpenAIs Whisper-modell er delt åpent og kan brukes som den er uten fintuning, blant annet for å transkribere norsk tale. Det er imidlertid også mulig å fintune slik at den blir bedre på et bestemt språk eller domene.[2] Whisper-modeller bruker ikke uttaleleksikon eller en egen språkmodell.

En interessant bieffekt av at Whisper er trent på undertekst, ikke-ordrette datasett, er at modellen også produserer transkripsjoner som ikke er ordrette. Eksempel (1) er tatt fra testsettet vi har brukt i denne rapporten. Vi indikerer her og i framtidige eksempler den ordrette, manuelle transkripsjonen med *gullstandard* og transkripsjonen fra modellen med navnet på modellen.[3]

(1)

> **lydfil:** Helgemorgen_021022_4432000_5111000_s530200_e533490.wav
> **gullstandard:** *Og det tenker jeg at det er det bør det være rom for.*
> **openai-whisper-large-bokmål:** *Det bør det være rom for.*

Lydfilen inneholder to omstarter av setninga, noe som er reflektert i gullstandarden: “Og det tenker jeg at” og “det er”. Den predikerte transkripsjonen har droppa disse omstartene og skriver kun “Det bør det være rom for.”

Denne effekten av Whisper-modeller er trolig nyttig i mange sammenhenger, men gjør samtidig testing krevende. Vi vil derfor komme tilbake til dette flere ganger i rapporten.

Norske talegjenkjenningssystemer

Her vil vi presentere de talegjenkjenningssystemene vi skal teste i denne rapporten. Disse inkluderer de viktigste kommersielle systemene for talegjenkjenning av spontan tale og systemer med åpen kildekode som er mye i bruk eller som etter vår vurdering representerer en bestemt talegjenkjenningsteknologi på en god måte. Vi tester ikke systemer som primært er ment for diktering, slik som Omilons Dragon-system. Vi har også sett bort fra en del mindre, kommersielle leverandører av talegjenkjenning, slik som Amberscript.

---

[2] I tillegg til å transkribere tale, kan Whisper også oversette tale til engelsk. Man kan for eksempel oversette norsk tale til engelsk tekst. Denne funksjonaliteten har vi ikke testa for denne rapporten.

[3] Når man omtaler datasett for testing og veileda læring av maskinlæringsmodeller, omtaler man gjerne den korrekte oppmerkinga som modellen skal lære seg å etterlikne, for *gullstandard*, til forskjell fra modellens predikerte oppmerking.

Flere av systemene har ulike alternative modeller. I disse tilfellene har vi valgt å teste noen få representative modeller. Modellene vi tester, refererer vi gjerne til med et kallenavn, f.eks. *openai-whisper-large* for OpenAIs store Whisper-modell. Disse kallenavnene brukes i koden vår, og grafer og tabeller vil ha disse kallenavnene. Vi synes det er nyttig å bruke de samme kallenavnene i teksten.

Nasjonalbibliotekets wav2vec2-modeller

I 2022 fintuna Nasjonalbibliotekets AI-lab ulike wav2vec2-modeller på norske datasett i samarbeid med Språkbanken. De forskjellige modellene benytter seg av ulike forhåndstrente modeller og er fintuna på ulike kombinasjoner av datasett. De la Rosa et al. (2023) presenterer de ulike modellene og testresultater for dem. Vi har testa modellene med best resultat for bokmål og nynorsk.

nb-wav2vec2-bokmål

Denne bokmåls-modellen bruker den forhåndstrente *voxrex*-modellen, en modell utvikla av Det kungliga bibliotek i Sverige, som er trent på 10 000 timer med lokalradiosendinger fra Sverige. Nasjonalbibliotekets AI-lab har fintuna denne modellen på bokmålsdelen av NPSC, en maksinoversatt versjon av nynorskdelen av NPSC, og NST.[4]

nb-wav2vec2-nynorsk

Denne nynorsk-modellen fintuner den flerspråklige, forhåndstrente XLS-R-modellen fra Meta på nynorskdata fra NPSC og en maskinoversatt versjon av bokmålsdelen av NPSC.

OpenAI Whisper

openai-whisper-large-bokmål

Dette er den offisielle, flerspråklige Whisper-modellen til OpenAI. Vi har testa *large*-versjonen, som trolig er den beste, men som må kjøres på kraftig datautstyr. Det fins også mindre versjoner av modellen som ikke krever like mye regnekraft, men som trolig er noe mindre nøyaktige. Treningsdataene for denne modellen er undertekst og lyd fra videoer som er høsta fra nettet. Det er ikke kjent hvor disse videoene er henta fra.

openai-whisper-large-nynorsk

Siden dette er en flerspråklig modell, er det samme modell som kjøres for *openai-whisper-large-bokmål* og *openai-whisper-large-nynorsk*, men vi har spesifisert at skriftspråket skal være henholdsvis bokmål eller nynorsk når vi har kjørt modellene.

[4] Ifølge de la Rosa et al. (2023) fins det en modell trent på samme datautvalg, men med en annen forhåndstrent modell (XLS-R), som har marginalt bedre resultater. I skrivende stund er denne modellen ikke tilgjengelig for åpen testing.

NB Whisper

Nasjonalbiblioteket er i skrivende stund (januar 2023) i ferd med å lansere en egen versjon av Whisper for norsk, NB Whisper. Betaversjoner har blitt distribuert siden sommeren 2023. NB Whisper kan regnes som en svært omfattende fintuning av OpenAI Whisper: Teamet bak NB Whisper har tatt utgangspunkt i vektene fra maskinlæringsmodellen til OpenAI Whisper, og har så trent videre på store mengder transkriberte, norske taledata. NB Whispers trening er omtrent like omfattende som OpenAI Whispers[5], og treningsdataene omfatter rundt 50 000 timer med transkribert tale. På grunn av den omfattende treningen, er det riktigere å regne NB Whisper som egne modeller med Whisper-arkitektur enn fintuna varianter av OpenAI Whisper.[6]

Treningsdataene for NB Whisper er NST, SSC, lyd og undertekst høsta fra NRK og lydbøker kobla med tekst fra Nasjonalbibliotekets samling.

NB Whisper kommer i tre ulike versjoner: Grunnversjonen av NB Whisper skal ha en transkripsjonsstil som ligger relativt tett opp til det som sies, uten nødvendigvis å transkribere helt ordrett. NB Whisper Semantic er trent opp til å gjengi meningsinnholdet, men gjerne med noe færre ord enn i talen. Denne er ment for møtereferater, undertekst o.l. Begge disse versjonene produserer store og små bokstaver og tegnsetting. NB Whisper Verbatim er kjørt gjennom noen ekstra fintuningssteg på NPSC, noe som gjør at den transkriberer ordrett. Denne produserer kun små bokstaver og ingen tegnsetting. NB Whisper Verbatim er ment for brukere som trenger ordrette transkripsjoner, f.eks. språkforskere, eller politi som skal transkribere avhør. Alle modellene støtter både nynorsk og bokmål. De kan også oversette fra norsk til engelsk og til en viss grad fra andre språk til norsk. Vi har ikke testa oversettelsesfunksjonaliteten. Som for OpenAI Whisper fins det modeller i ulike størrelser: large, medium, small og tiny. Vi tester bare large.

På tidspunktet denne rapporten skrives, er ikke de endelige versjonene av NB Whisper-modellene lansert ennå, og vi har derfor testa beta-modeller. Med unntak av noen mindre, tekniske justeringer som ikke skal påvirke transkripsjonene, vil de endelige modellene være like de vi har testa i dette prosjektet. Det er imidlertid i skrivende stund uklart om semantic–modellen kommer til å være med i den endelige versjonen.

Merk at siden NB Whisper er trent på lyd og undertekst fra NRK, er det en viss sannsynlighet for at den har blitt eksponert for noe av lydmaterialet i testsettet brukt her under treningen, siden testsettet også er laga på bakgrunn av lydopptak fra NRK TV. Det er umulig å si om dette har påvirka resultatene på noen måte, men vi anser det som lite sannsynlig at en eventuell påvirkning er spesielt stor. For det første vil lydmaterialet modellen eventuelt har blitt eksponert for, være en svært liten brøkdel av de totale treningsdataene fra NRK. For det andre har ikke modellen sett transkripsjonene fra

[5] Hovedmodellene til NB Whisper er trent 250 000 steg med en batch-størrelse på 1024, mens OpenAI er trent 1 million steg med en batch-størrelse på 256 ifølge Per Egil Kummervold, personlig kommunikasjon.

[6] Per Egil Kummervold, personlig kommunikasjon.

testsettet. Det faktum at modellen er trent på store mengder kringkastingsmateriale, vil imidlertid gjøre den bedre på å gjenkjenne kringkastingsmateriale. Det vil sannsynligvis påvirke resultatene i positiv retning når vi i denne undersøkelsen tester modellene på data fra NRK.

Google Cloud

Google Cloud, Googles skytjeneste, har tilbudt talegjenkjenning til bokmål i flere år. Fram til sommeren 2023 tilbød de kun ett system som vi vet relativt lite om, som vi i denne artikkelen kaller *gcloud-long*.[7] Vi vet ikke hva slags arkitektur som er brukt eller hva slags data modellen er trent på. Denne modellen støtter kun bokmål.

I 2023 lanserte Google en ny, flerspråklig talegjenkjenningsmodell som heter Universal Speech Model (USM). Bokmål er et av språkene modellen støtter. I Google Cloud kan man enten kjøre den gamle talegjenkjenningsmodellen eller USM, og vi har testa begge. USM har en proprietær arkitektur som minner noe om wav2vec2. Systemet bruker en forhåndstrent modell som er trent uveileda på 12 millioner timer med tale på ulike språk uten transkripsjon. Denne modellen er deretter fintuna på transkriberte datasett som delvis kommer fra lyd og undertekst fra Youtube og delvis fra åpne datasett (Zhang et al., 2023). Det er uklart hvilke data som er brukt for å trene den norske talegjenkjenningen til USM. Vi fikk særskilt teste-tilgang til USM før modellen ble lansert kommersielt, og det er denne tilgangen vi har brukt når vi har testa modellen.

Begge systemene til Google transkriberer ord for ord.

Microsoft Azure

Azure er Microsofts skyløsning og en av de store konkurrentene til Google Cloud. Azure tilbyr talegjenkjenning til mange språk, blant annet bokmål. Vi vet ingenting om hvilken teknologi som brukes i Azures talegjenkjenning og ei heller hvilke treningsdata de har brukt. Microsoft sa imidlertid i 2019[8] at de brukte transkriberte data fra Nasjonalbiblioteket til å utvikle sin talegjenkjenning, så det er sannsynlig at i det minste NST-datasettet inngår i treningsgrunnlaget. Azure ser ut til å transkribere relativt ordrett, men har tegnsetting og store og små bokstaver.

Transparens om medforfatteres involvering og om deling av resultater

For ordens skyld gjør vi oppmerksom på at Per Erik Solberg, medforfatter av denne rapporten, har ledet arbeidet med å lage SSC, som er del av treningsdataene til NB Whisper. Han har også rådgitt utviklingsteamet til NB Whisper på ulike vis og har rapportert om feil på

[7] Google tilbyr en versjon for lengre lydopptak, som de kaller *long*. De har også en for kortere opptak, som de kaller *short*. For å indikere at det er den første versjonen vi bruker, bruker vi også betegnelsen *long*. Long-modellen skal være den mest egna for den typen data vi har i testsettet. For sikkerthets skyld testa vi også short-modellen, som ga dårligere resultater enn long-modellen. Vi har ikke tatt med resultatene fra short-modellen i rapporten.

[8] https://www.tek.no/nyheter/nyhet/i/xPyjnQ/microsoft-lanserer-tale-til-tekst-paa-norsk

tidligere versjoner av modellene. Han har også vært involvert på tilsvarende måte i utviklingen av Nasjonalbibliotekets wav2vec2-modeller.

Aggregerte resultater av testing av tidligere versjoner av NB Whisper på testsettet har blitt delt med utviklerne av disse modellene, i tillegg til enkelte eksempler på transkripsjoner med feil, men utviklerne av NB Whisper har ikke fått tilgang til lydfiler fra testsettet og har ikke benyttet seg av testsettet i sine arbeidsløyper. Tilsvarende har utviklerne av USM-modellen til Google og nb-wav2vec2-bokmål fått tilgang til aggregerte resultater av modellene anvendt på testsettet og eksempler på transkripsjoner med kritiske feil. Vi har også delt lydfilene til noen få segmenter (totalt 6) med disse utviklerne for feilsøking av spesifikke feil. Disse lydfilene er tydelig merka og har blitt holdt til side når vi har testa modellene i denne rapporten. Testing på testsettet i en tidlig fase av prosjektet viste seg å belyse kilder til alvorlige mangler på systemer under utvikling som ikke så lett kunne bli oppdaga på andre vis. Vi syntes det var viktig å videreformidle slik informasjon med utviklerne for å sikre at norsk talegjenkjenning fungerer så godt som mulig, men har gjort vårt beste for å sikre at utviklerne ikke har anledning til å tilpasse sine modeller til dataene i testsettet.

Navn	Arkitektur	Utvikla av	Språk	Trenings-data	Tilgang	Hjemme-side	Referanse-artikkel	Dato for testing
nb-wav2vec2-bokmål	wav2vec2	NB	bokmål	NPSC, NST	åpen	lenke	de la Rosa et al. (2023)	22.6. 2023
nb-wav2vec2-nynorsk	wav2vec2	NB	nynorsk	NPSC, NST	åpen	lenke	de la Rosa et al. (2023)	28.11. 2023
openai-whisper-large-bokmål	Whisper	OpenAI	bokmål	data fra internett	åpen	lenke	Radford et al. (2023)	8.12. 2023
openai-whisper-large-nynorsk	Whisper	OpenAI	nynorsk	data fra internett	åpen	lenke	Radford et al. (2023)	8.12. 2023
nb-whisper-large-bokmål	Whisper	NB	bokmål	SSC, NST, data fra NRK, lydbøker	åpen	lenke		7.12. 2023
nb-whisper-large-nynorsk	Whisper	NB	nynorsk	SSC, NST, data fra NRK, lydbøker	åpen	lenke		7.12. 2023
nb-whisper-large-verbatim-bokmål	Whisper	NB	bokmål	SSC, NST, data fra NRK, lydbøker, NPSC	åpen	lenke		7.12. 2023
nb-whisper-large-verbatim-nynorsk	Whisper	NB	nynorsk	SSC, NST, data fra NRK, lydbøker, NPSC	åpen	lenke		7.12. 2023
nb-whisper large-semantic-bokmål	Whisper	NB	bokmål	SSC, NST, data fra NRK, lydbøker	åpen	lenke		7.12. 2023
nb-whisper-large-semantic-nynorsk	Whisper	NB	nynorsk	SSC, NST, data fra NRK, lydbøker	åpen	lenke		7.12. 2023
azure	ukjent	Microsoft	bokmål	ukjent	mot betaling	lenke		5.7. 2023
usm	usm	Google	bokmål	Youtube, åpne datasett	mot betaling	lenke	Zhang et al. (2023)	8.12. 2023
gcloud-long	ukjent	Google	bokmål	ukjent	mot betaling	lenke		28.11. 2023

Tabell 1: Talegjenkjenningsmodellene som er testa i prosjektet

4. Testsett, analysemetodikk og problemstillinger

Beskrivelse av testsettet

Testsettet inneholder 10 timer lydmateriale fra NRK sine radio- og tv-programmer. Vi har henta materiale fra NRK for å få et variert testsett med variasjon i programmer, temaer, dialekter, stemmer m.m. Testsettet består av 101 klipp fra forskjellige programmer med 409 talere, som hver er annotert med metadata om kjønn, dialekt og opptaksforhold. Lydklippene er mellom 1 og 15 minutter lange, og de er blitt segmentert i enkeltsetninger og transkribert. Segmentene er maksimalt 30 sekunder lange, og lange setninger som overstiger 30 sekunder er delt opp i flere segmenter. Det er totalt 7261 segmenter i testsettet, og hele testsettet er transkribert på både bokmål og nynorsk.

Lydmaterialet

Testsettet er utvikla for å teste talegjenkjenningssystemer som er ment for dokumentasjon, teksting, referater o.l. og er ikke mynta på testing av dikteringssystemer eller talestyring av enheter. Tanken har vært å samle inn den typen lydmateriale som omgir oss i hverdagen og som det vil være naturlig å bruke talegjenkjenning for å transkribere automatisk. Dette inkluderer f.eks. radio- og tv-programmer, podkaster, foredrag, møter, intervjusituasjoner osv.

Testsettet skal i hovedsak brukes til å teste spontant, ikke-normert talespråk der personer snakker fritt og har et muntlig språk. I og med at norsk talespråk har stor dialektvariasjon, er det en prioritet å teste hvor gode talegjenkjenningssystemene er til å kjenne igjen forskjellige dialekter. I tillegg skal testsettet teste noen former for ikke-spontan tale. Her mener vi manusbasert tale som er ment for muntlig formidling, som f.eks. talen til et nyhetsanker, eller fortellerstemmen i en radiodokumentar. Det er derfor et spenn av forskjellig type tale fra helt dagligdagse samtaler mellom flere personer, til manuslest tale av én person. Det har ikke vært relevant å ta med lyd der noen leser en skriftlig tekst høyt, eller dramatisert materiale.

Det er viktig å ha i mente at selv om ikke alle programmene har manus vil det ofte være en viss grad av forberedelse, og selv i de dagligdagse samtalene vil personene likevel være bevisste på at de er med i et opptak. Det i seg selv kan gjøre at man snakker annerledes enn man vanligvis gjør, for eksempel ved å snakke tydeligere, snakke i mer fullstendige setninger, eller unngå visse ord som slang eller banneord.

Testsettet har også variasjon i opptakskvalitet og bakgrunnsstøy for å kunne teste hvor gode systemene er til å kjenne igjen tale under forskjellige opptaksforhold. Testsettet inneholder blant annet studioopptak med god kvalitet, opptak med bakgrunnsmusikk, trafikkstøy, ekko og lydeffekter, for å nevne noe.

I tillegg har vi sørga for en spredning i alder blant talerne. Det er ikke alltid mulig å si hvor gammel en person er ut ifra stemmen. Unge personer kan høres eldre ut, og eldre personer kan høres yngre ut. Av den grunn gir det heller ikke mening å samle inn en viss mengde tale

fra alle aldre, men vi har valgt noe materiale med yngre og eldre talere basert på tema (ungdomsprogrammer, dokumentar om eldre, nyhetssaker om unge/eldre o.l.). Talerne spenner fra ungdommer rundt 15 år til eldre personer rundt 80 år. Det er ikke med barnestemmer i testsettet fordi det på tidspunktet da testsettet ble laget (2022) ikke fantes treningsressurser med barnestemmer, og norske talegjenkjenningssystemer var ikke utviklet for å håndtere barnestemmer. Av samme grunn er det heller ikke med talere med utenlandsk aksent i testsettet. Alle talerne har morsmålskompetanse i norsk og snakker en norsk dialekt.

Transkripsjon

Transkripsjonene er ortografiske og følger enten bokmåls- eller nynorsknormen. Testsettet ble først transkribert automatisk på bokmål. De automatiske transkripsjonene ble deretter retta av lingvister i Språkbanken i henhold til gitte retningslinjer, og så korrekturlest. De ferdige bokmålstranskripsjonene ble oversatt automatisk til nynorsk med Apertium før disse også ble korrekturlest.

Transkripsjonene er en skriftlig representasjon av det som blir sagt i lydmaterialet. Setningene begynner med stor bokstav og har tegnsetting som punktum og komma. Tall er skrevet med siffer, ikke bokstaver, og vi har tillatt bruk av forkortelser, måleenheter og noen symboler, som prosenttegn.

Samtidig ligger transkripsjonene nært opp til det talerne faktisk sier. Forskjellige systemer vil ha forskjellig praksis for hvor mye de retter eller forenkler språket i den automatiske transkripsjonen, og noen av de Whisper-baserte modellene vil gi en undertekstlignende transkripsjon som forenkler utsagnet i svært stor grad. Vi har derfor måttet ta noen valg rundt hva og hvor mye vi vil rette på språket i transkripsjonene. Vi har valgt å ikke gjøre for mange antakelser, og transkripsjonene er nesten en ord-for-ord-gjengivelse av talen. Vi har fjerna overflødige repetisjoner av enkeltord og vi har retta noen typer språklige feil, som feil bøying av ord. Utover dette har vi ikke endra eller retta språket for å skape syntaktisk velforma setninger. I muntlig språk snakker vi ikke alltid i grammatisk riktige setninger, og setningene i testsettet er derfor heller ikke alltid grammatisk riktige.

Der det er valgfrihet innenfor skriftnormen har vi i bokmålstranskripsjonen valgt en talenær transkripsjon. For eksempel er det valgfritt med a- og en-endelse på hunkjønnsord i bokmål (*boka/boken*), og vi har transkribert med den formen som ligger nærmest det taleren sa i hvert tilfelle. I nynorsktranskripsjonen har vi valgt å transkribere med standardformene Apertium gir, både for å lette arbeidet med oversetting og korrektur, men også for å følge det som anses som konvensjonene innenfor målforma. Det vil f.eks. si at nynorsktranskripsjonene har a-infinitiver, ikke e-infinitiver, og at verbet *å bli* bøyes *blir – vart – har vorte*, ikke *blir – blei – har blitt*. De forskjellige talegjenkjenningssystemene vil ta forskjellige valg innenfor bokmåls- og nynorsknormen. Noen vil bruke konsekvente former i transkripsjonen, mens andre vil gjengi talen mer nøyaktig. Det er derfor fordeler og ulemper ved begge fremgangsmåtene, og både bokmåls- og nynorsktranskripsjonene vil ha uoverensstemmelser med den automatiske transkripsjonen.

Kjønn, dialektkategorier og opptaksforhold

Testsettet har metadata om kjønn, dialektområde, finkorna dialekt og opptaksforhold for hver taler.

Kjønn

Testsettet har to kjønnskategorier: kvinne (f) og mann (m). Hver av kategoriene utgjør halvparten av testsettet (f 50,7 %, m 49,3 %), og de to kategoriene rommer et bredt spekter av forskjellige stemmer.

Dialektområde

Det er fem overordna dialektområder i testsettet:

- Østlandsk (e) (Innlandet, Oslo, Viken, Vestfold og Telemark, Agder)
- Sørvestlandsk (sw) (Rogaland)
- Vestlandsk (w) (Vestland, Møre og Romsdal)
- Trøndersk (t) (Trøndelag)
- Nordnorsk (n) (Nordland, Troms og Finnmark)

Det er vanskelig å få til en absolutt femdeling av testsettet med to timer per dialektområde, men hver av områdene er godt representert med minst 1 time og 15 minutter. Østlandsk er området med størst representasjon.

Finkorna dialekt

De fem overordna dialektområdene er også brutt ned i mer spesifikke dialekter, og alle talerne er blitt annotert med informasjon om den mer finkorna dialekta i tillegg til dialektområdet.

- Østlandet
    - Oslo (Oslo og oslonære dialekter. Dette omfatter områdene Oslo, Akershus, Vestfold og flatbygdene i Telemark og Buskerud).
    - Innlandet (Innlandet bortsett fra Valdres og Nord-Gudbrandsdalen).
    - Midtlandsk (Dalstrøkene og fjelltraktene midt i landet: deler av Telemark, Numedal, Hallingdal, Valdres, Nord-Gudbrandsdal).
    - Agder
    - Østfold
- Sørvestlandet
    - Rogaland
- Vestlandet
    - Hordaland u/Bergen (Hordaland utenom Bergen).
    - Bergen (Bergen og bergensnære dialekter).
    - Sogn og Fjordane
    - Sunnmøre
    - Romsdal + Nordmøre
- Trøndelag
    - Trøndelag

- Nord-Norge
    - Nordland
    - Troms
    - Finnmark

Alle disse dialektene er representert i testsettet, og vi har blant annet hentet materiale fra lokalnyheter, men det har ikke vært mulig å sørge for en jevn fordeling av dialektene innenfor hvert område fordi det ville blitt en for tidkrevende oppgave.

Den finkorna dialektinndelinga reflekterer de geografiske områdene der dialektene har nok likhetstrekk til at det gir mening å anse dem som ett dialektalt område. Inndelinga støtter seg dessuten på tradisjonelle inndelinger man finner i litteraturen, og både fonologi og formverk er tatt med i beregningen for å avgjøre hvorvidt dialekter har store likhetstrekk eller ei. Det har ikke gitt mening å dele dialektene inn i enda mindre områder fordi de geografiske områdene da blir så små at testsettet ikke inneholder talere fra alle stedene.

Opptaksforhold

Opptaksforholdene er delt inn i tre kategorier, og talerne er annotert med verdi 1, 2 eller 3:

1. Studioopptak og andre innendørs og utendørs opptak i stille omgivelser
2. Innendørs og utendørs opptak med støy eller bakgrunnsmusikk
3. Opptak med telefonkvalitet

Det er 409 talere i de forskjellige lydklippene i testsettet, men noen personer opptrer i flere klipp og det er kun 360 unike personer. Vi skiller mellom taler og person fordi samme person f.eks. kan snakke i støyende omgivelser i ett klipp, men i studio i et annet. Talernivået kan med andre ord beskrives som alle segmentene til samme person innenfor samme klipp. Det betyr også at det i realiteten er segmentene til taleren som blir annotert med opptaksforhold. Siden opptaksforholdene som regel er de samme for samme taler gjennom hele klippet, har vi valgt å annotere opptaksforhold på talernivå og alle talerens segmenter får én og samme verdi.

Vi har fulgt tommelfingerregelen at svak bakgrunnslyd, som suselyd fra vind eller trafikk, blir annotert med 1 så lenge opptaket på det hele oppfattes som stille. Opptaket annoteres med 2 når det er tydelige lyder eller musikk i bakgrunnen, selv om det ikke er så støyende at det er vanskelig å oppfatte talen. Kategori 2 har dermed et spenn av støy fra lyder som ikke forstyrrer opptaket i særlig grad til høylytt støy som forstyrrer opptaket og talen. Alle opptakene har også større eller mindre grad av ekko avhengig av opptaksomgivelsene. Stille opptak med normal mengde ekko anses som stille og er annotert med 1, mens opptak som har så mye ekko at lyden oppfattes som støyende er annotert med 2.

Kvalitetsmål for talegjenkjenning

Talegjenkjenningssystemer trenes på datasett med lydopptak og korrekte transkripsjoner av disse lydopptakene. Lydopptak og korrekte transkripsjoner brukes også for å teste talegjenkjenningssystemer. De transkriberte lydopptakene i et testsett for talegjenkjenning er

som regel på opp til 30 sekunder og kalles *segmenter*. I vårt testsett er hvert segment lydopptak og transkripsjon av én setning.

Når man skal teste et talegjenkjenningssystem, får man talegjenkjenningssystemet til å transkribere hvert segment. Så sammenlikner man den korrekte transkripsjonen av segmentene, ofte kalt *gullstandarden*, med den automatisk genererte transkripsjonen, kalt *prediksjonen*, og regner ut hvor forskjellige prediksjonene er fra gullstandarden. Det er flere måter å regne ut denne forskjellen på, som alle har sine fordeler og ulemper. De to vanligste er *ordfeilrate* (*Word Error Rate*, forkorta i det følgende som *WER*) og *tegnfeilrate* (*Character Error Rate*, forkorta i det følgende som *CER*). Vi bruker også et tredje mål, *SemDist*.

Ordfeilrate (WER)

WER er andelen ord i den automatiske transkripsjonen som er lagt til, bytta ut eller mangler, sammenlikna med gullstandarden. For eksempel er WER i eksempel (2) 18,18% fordi talegjenkjenningssystemet har splitta opp «hardtslående» i to ord (ett ord er bytta ut + ett ord er lagt til / 11 ord * 100).[9] De orda som er forskjellig, er utheva i de to transkripsjonene.
(2)
lydfil:
Filmpolitiets_podkast_To_ferske_krimtips_fra_Los_Angeles_120522_2130000_2683000_s383070_e385579.wav
gullstandard: *jeg regner med at regner med at den er ganske **hardtslående***
usm: *jeg regner med at regner med at den er ganske **hardt slående***

Ordfeilrate er det mest brukte målet på kvaliteten til talegjenkjenningssystemer. Microsoft sier i denne artikkelen at et system som gir en ordfeilrate på 5-10% er av høy kvalitet og klar til å tas i bruk. En ordfeilrate på 20% er akseptabel, og en ordfeilrate på over 30% er å regne som lav kvalitet. Disse vurderingene er et nyttig grunnlag for diskusjon, men som vi skal se, samsvarer ikke alltid ordfeilraten med opplevd kvalitet. Ordfeilraten ved manuell transkripsjon er på mellom 4 og 9% (Stolcke & Droppo, 2017).

Tegnfeilrate (CER)

CER regnes ut på samme måten som WER, men istedenfor å se på andelen ord som er lagt til, bytta ut eller mangler, ser man på andelen tegn (bokstaver og mellomrom) som er lagt til, bytta ut eller mangler. CER for eksempel (2) er 1,72%. Merk at CER er mye lavere enn WER, som var på 18,8%: Den eneste forskjellen på tegnnivå mellom gullstandarden og prediksjonen er at et mellomrom er lagt til i prediksjonen, mens det feilaktige mellomrommet fører til større forskjell når man måler på ordnivå.

WER og CER er standardmål for kvaliteten på talegjenkjenning, men WER er nok aller mest brukt. Fordelene med disse målene er at de er svært enkle å regne ut og forstå, og de gir

[9] Før vi regner ut WER og CER i våre analyser, fjerner vi tegnsetting og gjør alle bokstaver små. Vanligvis vil vi imidlertid oppgi gullstandardtranskripsjon og predikert transkripsjon med store og små bokstaver og tegnsetting. I dette eksempelet har vi ikke gjort det for at utregninga skal være mer transparent.

ofte et godt bilde av kvaliteten på systemer som transkriberer ordrett. WER er noen ganger litt for rigid: La oss si at et system skriver *betydningen*, mens gullstandarden har *betydninga*. For en norsktalende vil dette kanskje ikke bli regna som en feil i det hele tatt, eller hvertfall en ganske liten feil, men for utregningen av WER vil dette telle som én feil like mye som om systemet hadde skrevet et helt urelatert ord. CER kan i slike tilfeller gi et bedre bilde, siden kun bokstavene *e, n* og *a* er ulike mellom de to formene.

Et annet problem med WER og CER er at de ikke tar hensyn til ords betydning. Dersom en prediksjon kun mangler ordet *ikke* mens gullstandarden har dette ordet, vil WER være ganske lav, siden det kun er en forskjell på ett ord. Men å fjerne *ikke* vil som oftest føre til at setninga betyr noe helt annet.

Problemene med WER og CER er enda større for systemer som ikke transkriberer ordrett. Det ligger et premiss til grunn for disse kvalitetsmålene at en god transkripsjon er en ordrett transkripsjon. Whisper-baserte systemer kan imidlertid produsere transkripsjoner som oppleves som gode, men som ikke er ordrette, og dermed får høy WER og CER. Det er tilfelle i eksempel (1), gjentatt i (3):

(3)

lydfil: Helgemorgen_021022_4432000_5111000_s530200_e533490.wav
gullstandard: *Og det tenker jeg at det er det bør det være rom for.*
openai-whisper-large-bokmål: *Det bør det være rom for.*

Her er WER 53,85%, som er å regne som et svært dårlig resultat. Imidlertid vil den predikerte transkripsjonen fungere godt i mange sammenhenger, for eksempel i undertekst, og vil kanskje være vel så nyttig som den ordrette gullstandard-transkripsjonen til slike formål.

På grunn av dette problemet vil Whisper-baserte systemer typisk få en langt dårligere WER og CER enn systemer som transkriberer ordrett, og WER og CER korresponderer dårlig med den opplevde kvaliteten til disse systemene.

SemDist

For å komplementere bildet noe, har vi regna ut et tredje kvalitetsmål i tillegg til WER og CER: SemDist. SemDist står for *semantic distance* og ble foreslått av forskere hos Meta i Kim et al. (2021) som et supplement til WER og CER for å bøte på noen av begrensningene til disse kvalitetsmålene. SemDist er på ingen måte like transparent og lett å regne ut som WER og CER. For å regne ut SemDist, kjører man en maskinlæringsmodell på gullstandarden og prediksjonen som regner ut en *setningsembedding* for hver av disse transkripsjonene. En setningsembedding er en vektor, en sekvens med tall, som representerer betydninga til setninga ifølge maskinlæringsmodellen. SemDist er et mål på avstanden mellom embeddingen til gullstandarden og prediksjonen. Jo lavere SemDist, desto likere er betydninga til de to transkripsjonene. For å regne ut SemDist har vi brukt modellen *nb-sbert-base* utvikla av AI-laben ved Nasjonalbiblioteket.[10] I motsetning til WER og CER har ikke SemDist noen absolutt fortolkning. En SemDist på 0,005 sier ikke isolert

[10] https://huggingface.co/NbAiLab/nb-sbert-base

sett om to transkripsjoner er like eller forskjellige; det sier kun noe om hvor like disse transkripsjonene er sammenlikna med andre SemDist-utregninger fra samme modell. Så to transkripsjoner med en SemDist på 0,005 kan man anta er likere i betydning enn to fra samme modell med SemDist på 0,05 og mer ulike i betydning enn to med SemDist på 0,0005.

SemDist har et høyere abstraksjonsnivå enn WER og CER: Den sier noe om hvor like to transkripsjoner er i betydning, uavhengig av hvilke ord eller bokstaver som brukes. To transkripsjoner som bruker ganske forskjellige ord, kan dermed få relativt lav SemDist dersom de betyr det samme. Dersom et system ikke transkriberer ordrett, men betydninga av de predikerte transkripsjonene ofte er like betydninga til gullstandardtranskripsjonene, vil SemDist være lav. Dermed kan SemDist være et nyttig mål på kvaliteten på for eksempel Whisper-baserte systemer, siden disse ikke alltid transkriberer ordrett.

En ulempe med SemDist er at utregninga ikke er transparent. Den bruker en svært kompleks maskinlæringsmodell til utregninga, og det er umulig å vite presis hvorfor den modellen produserer akkurat de vektorene den gjør. Denne maskinlæringsmodellen kan forledes av skjevheter i tekstdataene den er trent på, og det kan påvirke SemDist, uten at det er så lett å oppdage.

Ved å bruke SemDist som kvalitetsmål antar man at en predikert transkripsjon som betyr det samme som gullstandarden, er en god transkripsjon, uavhengig av hvilke ord som brukes. Dette er en antakelse som bare delvis er sann. Skal en transkripsjon brukes i for eksempel avhør, er det trolig viktig at den i så stor grad som mulig bruker akkurat de samme ordene som taleren sier. For undertekst er det nok større fleksibilitet, men det er sannsynlig at det også der vil oppfattes som rart om ordene avviker mye fra det som faktisk sies. Til syvende og sist er det kun ved å spørre brukere man får vite hvor gode og nyttige transkripsjoner oppleves å være.

Problemstillinger

Som nevnt i kapittel 1, er hovedmålene for testinga:

1. Evaluere status for norsk talegjenkjenning
2. Finne ut hva slags feil norske talegjenkjenningssystemer gjør og hva Språkbanken skal bruke tid og ressurser på videre
3. Vise forbedringene i norsk talegjenkjenning over tid

For å nå disse målene, har vi formulert seks problemstillinger, to for hvert mål:

- 1 a) Ved bruk av tilgjengelige kvalitetsmål for talegjenkjenning, hvilke resultater får vi på vårt testsett?
- 1 b) For hvilke modeller fungerer disse kvalitetsmålene godt, og for hvilke kommer det til kort på testsettet?
- 2 a) Hvor gode er systemene til å håndtere ulike dialekter, kjønn, opptaksforhold og overlappende tale?

- 2 b) Hvilke andre faktorer påvirker kvaliteten på transkripsjonene til de forskjellige systemene?
- 3 a) På hvilken måte har ny modellarkitektur påvirket kvaliteten på norsk talegjenkjenning?
- 3 b) På hvilken måte har data fra Språkbanken påvirket kvaliteten på norsk talegjenkjenning?

Hadde kvalitetsmål som WER og CER vært gode kvalitetsmål for alle modellene, hadde vi kunnet klart oss uten problemstilling 1 b. Siden disse målene er mindre egna for ikke-ordrette modeller, er det imidlertid relevant å drøfte hvor gode målene er og hvor de kommer til kort.

Analysemetodikk

Vi har kjørt alle segmentene i testsettet gjennom alle talegjenkjenningssystemene som skal testes, og har regna ut WER, CER og SemDist for hvert segment. Med utgangspunkt i dette har vi regna ut gjennomsnittlige verdier for hele testsettet og for deler av testsettet, for eksempel for hver dialekt.[11] Utregningene av WER og CER har blitt gjort på transkripsjoner som er noe modifisert: Tegnsetting har blitt fjerna og alle bokstaver er gjort små. Vi har også fjerna oppmerking av nølelyder som *eh* og *ehm* fra transkripsjonene før måling av WER og CER. Dette har vi gjort for å isolere de språklige forskjellene, som er viktigst for denne analysen. SemDist har vi imidlertid regna ut på umodifisert tekst, siden SemDist antakelig har nytte av tegnsetting og store og små bokstaver i utregninga.

I tillegg til disse aggregerte analysene har vi også spilt av lydfilene til en del segmenter og gjort en kvalitativ analyse av transkripsjonene. Vi har også brukt noen andre analysemetoder der det er relevant, noe vi kommer tilbake til.

5. Analyse av ASR-systemer

Vi har nå kommet til selve analysen av talegjenkjenningssystemene. Dette kapittelet er delt inn i paragrafer for hvert av hovedmålene våre og så underparagrafer for problemstillingene.

[11] WER og CER er sensitive for segmentlengde: En feil på ett ord vil ha mye større innflytelse på WER i en kort setning enn en lang setning, og tilsvarende for CER. Siden CER og WER for en kort og en lang setning ikke kan sammenliknes direkte, blir det ikke riktig å regne ut WER og CER ved å summere opp WER/CER-verdiene for alle segmentene og så dele på antall segmenter. I stedet har vi for hvert segment multiplisert WER-ratioen med antall ord i gullstandarden og CER-ratioen med antall tegn i gullstandarden for å få antallet ordfeil/tegnfeil. Så har vi regna ut prosenten av ordfeil/tegnfeil i det totale antallet ord/tegn i datasettet.

Status for talegjenkjenning på norsk

1 a) Ved bruk av tilgjengelige kvalitetsmål for talegjenkjenning, hvilke resultater får vi på vårt testsett?

Bokmål

Tabell 2 viser ordfeilraten på testsettet for de forskjellige modellene på bokmål. Kolonnen *WER* gir den gjennomsnittlige ordfeilraten på hele testsettet, mens *WER 25%*, *WER 50%* og *WER 75%* rapporterer den gjennomsnittlige ordfeilraten i de beste 25, 50 og 75 prosentene av dataene. De beste resultatene i hver kolonne er skrevet med fet skrift. Modellene er sortert etter WER fra lavest til høyest.

Modell	**WER**	**WER 25%**	**WER 50%**	**WER 75%**
nb-whisper-large-verbatim-bokmål	**11,89%**	**0,00%**	**2,01%**	**6,58%**
nb-whisper-large-bokmål	18,71%	**0,00%**	3,81%	10,11%
usm	20,84%	0,98%	6,81%	12,67%
nb-wav2vec2-bokmål	22,32%	2,47%	9,31%	16,68%
azure	24,63%	1,51%	8,69%	15,57%
openai-whisper-large-bokmål	30,17%	5,79%	14,69%	22,61%
gcloud-long	30,40%	7,00%	14,34%	23,68%
nb-whisper-large-semantic-bokmål	31,53%	**0,00%**	10,77%	20,90%

Tabell 2: Ordfeilrate på bokmål

Modellen med lavest ordfeilrate på hele datasettet er nb-whisper-large-verbatim-bokmål. Som alle NB-Whisper-modellene er denne modellen trent på mer enn 50 000 timer. I tillegg er den optimalisert for ordrett transkripsjon, som WER måler. Mer overraskende er det kanskje at nb-whisper-large-bokmål kommer på andreplass. I likhet med openai-whisper-large-bokmål og nb-whisper-large-semantic-bokmål transkriberer ikke denne helt ordrett. Likevel får den bedre WER enn modeller som skal skrive ordrett, som usm og nb-wav2vec2-bokmål, noe som både indikerer at den gjør relativt små modifikasjoner når den ikke transkriberer ordrett, og at transkripsjonene har høy kvalitet. nb-whisper-large-semantic-bokmål har imidlertid den høyeste ordfeilraten av alle modellene, så der er det relativt stort avvik mellom gullstandarden og de automatiske transkripsjonene.

OpenAIs Whisper-modell, openai-whisper-large-bokmål, har den tredje høyeste ordfeilraten av alle modellene. Denne modellen er mye brukt for tida, for eksempel av UiO og Schibsted, [12] og brukere er gjerne godt fornøyd med den, noe som indikerer at WER ikke nødvendigvis samsvarer med opplevd kvalitet.

Ingen av de kommersielle modellene, usm, azure og gcloud-long, er blant de beste modellene målt i ordfeilrate. Da vi målte en beta-versjon av Googles USM-modell i juni, var WER på i underkant av 17%, betydelig under 20,84%, prosenten vi fikk da vi testa på nytt i desember.

Gjennomsnittlig WER for de 25%, 50% og 75% beste dataene sier noe om spredningen av feil. Det er interessant å observere at alle NB-Whisper-modellene har en WER på 0% i de 25% beste dataene, som indikerer at disse modellene ofte transkriberer helt korrekt. Til sammenlikning har openai-whisper-large-bokmål, som bygger på samme teknologi, 5,79% WER i dette utvalget, den nest høyeste verdien etter gcloud-long.

Tabell 3 rapporterer gjennomsnittlig tegnfeilrate på hele datasettet.

Modell	**CER**
nb-whisper-large-verbatim-bokmål	**6,50%**
nb-wav2vec2-bokmål	11,95%
usm	12,89%
nb-whisper-large-bokmål	13,30%
azure	16,05%
openai-whisper-large-bokmål	18,31%
gcloud-long	22,28%
nb-whisper-large-semantic-bokmål	24,68%

Tabell 3: Gjennomsnittlig tegnfeilrate

Rangeringen av modellene er den samme som for WER med ett unntak: nb-wav2vec2-bokmål har det nest laveste resultatet i stedet for nb-whisper-large-bokmål. Wav2vec2-modeller har en tendens til å transkribere ord tett opp til slik de uttales. Om de blir eksponert for ord de har sett lite i treningsmaterialet sitt eller for dialekter de ikke er gode på, kan de av og til transkribere nært opp til slik ordet blir uttalt, selv når det bryter med offisiell ortografi. For eksempel kan *uka* bli transkribert som *veka* for talere som har den dialektformen. For ordfeilraten vil slike skrivefeil telle som én feil på lik linje med andre feiltranskriberte ord. Tegnfeilraten vil imidlertid ta hensyn til at det er overlappende bokstavsekvenser i de to ordene. Det er sannsynlig at dette er årsaken til at

[12] https://www.uio.no/tjenester/it/aktuelt/om-it/2023/autotekst-1.html
https://www.kode24.no/artikkel/sann-lagde-vg-utviklerne-jojo-med-openai-losning/78756479

wav2vec-modellen gjør det relativt bedre for CER enn for WER. På tross av disse forskjellene er WER og CER relativt like typer mål og vil ofte vise de samme mønstrene. I det følgende vil vi ofte kun rapportere WER.

Tabell 4 presenterer gjennomsnittlig SemDist på hele datasettet. Modellene er rangert fra lavest til høyest SemDist.

Modell	**SemDist**
nb-whisper-large-bokmål	**0,09**
nb-whisper-large-semantic-bokmål	0,11
nb-whisper-large-verbatim-bokmål	0,12
azure	0,17
openai-whisper-large-bokmål	0,20
nb-wav2vec2-bokmål	0,21
usm	0,26
gcloud-long	0,38

Tabell 4: Gjennomsnittlig SemDist

Hvis vi sammenlikner tabell 4 med tabell 2 og 3, ser vi at nb-whisper-large-semantic-bokmål går fra å ha den høyeste WER- og CER-verdien til å ha den nest laveste SemDist-verdien. Det viser at selv om nb-whisper-large-semantic-bokmål ikke alltid transkriberer ordrett det som sies, er transkripsjonene fra denne modellen betydningsmessig like gullstandarden. Vi ser et motsatt mønster med usm, den nye talegjenkjenningsmodellen fra Google. Selv om usm har relativt lave verdier for WER og CER, har modellen den nest høyeste SemDist-verdien. Microsoft Azures modell har betydelig lavere SemDist enn usm. Gcloud-long har betydelig høyere SemDist enn de andre modellene, som viser at den har en relativt dårligere gjengivelse av meningsinnholdet i segmentene i testsettet.

Nynorsk

Tabell 5 gir gjennomsnittlig WER for nynorsk. Verdiene for de beste modellene er en del høyere enn WER-resultatene for bokmål rapportert i tabell 2. Tallene er imidlertid ikke helt sammenliknbare, da nynorsk har en annen type normvariasjon enn bokmål har. Blant annet tillater nynorsk infinitivsformer som ender på *-a* og på *-e*. Gullstandarden har konsekvent a-infinitiv, mens de automatiske transkripsjonene enten kan ha a-infinitiv eller e-infinitiv. Hvis for eksempel gullstandarden har infinitiven *å elska*, mens den automatiske transkripsjonen har *å elske*, vil dette bli regna som en feil, selv om begge formene er tillatt.

Modell	WER	WER 25%	WER 50%	WER 75%
nb-whisper-large-verbatim-nynorsk	**21,64%**	**1,76%**	**9,52%**	**15,21%**
nb-whisper-large-nynorsk	29,54%	2,22%	11,17%	18,93%
nb-whisper-large-semantic-nynorsk	36,18%	3,69%	15,13%	25,85%
nb-wav2vec2-nynorsk	39,99%	16,21%	25,24%	33,37%
openai-whisper-large-nynorsk	53,52%	25,56%	34,53%	41,82%

Tabell 5: Gjennomsnittlig ordfeilrate - nynorsk

Vi ser at nb-whisper-large-verbatim har lavest ordfeilrate på nynorsk som på bokmål, og at de tre NB Whisper-modellene har de tre øverste plassene. Openai-whisper-large-nynorsk har betydelig dårligere WER enn de andre systemene som støtter nynorsk.

I tabell 6 finner vi den gjennomsnittlige tegnfeilraten for systemene som støtter nynorsk.

Modell	CER
nb-whisper-large-verbatim-nynorsk	**12,45%**
nb-whisper-large-nynorsk	20,98%
nb-wav2vec2-nynorsk	21,62%
nb-whisper-large-semantic-nynorsk	28,06%
openai-whisper-large-nynorsk	47,10%

Tabell 6: Gjennomsnittlig tegnfeilrate - nynorsk

Som på bokmål ser vi at wav2vec2-modellen gjør det relativt bedre med dette kvalitetsmålet enn med ordfeilrate, antakelig fordi modellen produserer skrivefeil som ligger nært opp til den normerte formen.

Til slutt, la oss se på gjennomsnittlig SemDist for nynorsktranskripsjonene i tabell 7:

Modell	CER
nb-whisper-large-nynorsk	**0,11**
nb-whisper-large-semantic-nynorsk	**0,11**
nb-whisper-large-verbatim-nynorsk	0,13
nb-wav2vec2-nynorsk	0,28
openai-whisper-large-nynorsk	0,28

Tabell 7: Gjennomsnittlig tegnfeilrate - nynorsk

NB Whisper-modellene har lavest SemDist, mens wav2vec2-modellen og OpenAI Whisper har relativt høye SemDist-verdier sammenlikna med NB Whisper. Dette indikerer at NB Whisper gjengir meningsinnholdet bedre enn de to andre systemene. Merk imidlertid at dette målet ikke sier noe om hvor korrekt nynorsken er i transkripsjonene. SemDist-verdien for en setning på bokmål målt mot den samme setninga på nynorsk vil trolig være lav, fordi setningene betyr det samme.

1 b) For hvilke modeller fungerer disse kvalitetsmålene godt, og for hvilke kommer det til kort på testsettet?

WER og CER sier hvor mange ord eller tegn som er ulike i de automatiske transkripsjonene og gullstandarden. SemDist sier noe om hvor lik betydning de automatiske transkripsjonene og gullstandarden har, uavhengig av hvilke ord som er brukt. Det er med andre ord ganske ulike kvalitetsmål. For noen modeller samsvarer likevel SemDist godt med WER og CER. nb-whisper-large-verbatim-bokmål har relativt lave verdier med alle målene, og gcloud-long har relativt høye verdier med alle målene. For disse modellene ser alle målene ut til å bekrefte de samme tendensene, og i den forstand fungerer målene godt for disse modellene. Det er likevel verdt å diskutere i hvilken grad disse kvalitetsmålene svarer til opplevd kvalitet. Det spørsmålet skal vi komme tilbake til.

Noen modeller har imidlertid høy WER og CER og lav SemDist. Det gjelder særlig nb-whisper-large-semantic-bokmål, som har en av de laveste SemDist-verdiene, men har høy WER og CER. Dette har trolig å gjøre med at modellen ikke transkriberer ordrett i alle tilfeller. Motsatt har usm relativt lav WER og CER og relativt høy SemDist. Sier denne forskjellen noe meningsfullt om kvaliteten på disse modellene? For å undersøke dette nærmere har vi inspisert et utvalg segmenter fra hver av disse to modellene. Utvalget er gjort slik: For nb-whisper-large-semantic-bokmål har vi først identifisert de 30% av segmentene med høyest WER. I dette utvalget har vi så identifisert de 50% med lavest SemDist og har inspisert 10 tilfeldig utvalgte setninger fra dette subsettet. Disse setningene har altså relativt høy WER, men relativt lav SemDist. Dersom SemDist gir en god indikasjon på hvor lik den automatiske transkripsjonen er det som ble sagt, uavhengig av ordene som blir brukt, bør vi få setninger som betyr det samme som gullstandarden, men som bruker noe andre ord. For usm har vi gjort det motsatte: Vi har plukka ut de 30% av segmenter med høyest SemDist og så har vi valgt ut 10 tilfeldige setninger fra de 50% med lavest WER.

Når vi ser på dette utvalget med transkripsjonene fra nb-whisper-large-semantic-bokmål, er 8 av 10 tilfeller eksempler på ikke-ordrett transkripsjon som er kortere enn gullstandarden. De to gjenværende tilfellene er veldig korte segmenter. I korte segmenter vil få feil gjøre store utslag på WER fordi antall ord er del av utregninga for WER. I de 8 forkorta transkripsjonene har den automatiske transkripsjonen en mer skriftlig stil enn gullstandarden. I flere av eksemplene gjengis meningen relativt godt, som i setningene (4) og (5):

(4)

lydfil:
To_i_campingstol_Rita_Karin_Nyland_og_Roger_Sevrin_Bruland_210720_497000_1087000_s554580_e562600.wav
gullstandard: ***Og han*** *kompisen hans i Milano,* ***han dreiv jo og*** *var på Tinder-dater, og han dreiv jo på med forskjellig.*
nb-whisper-large-semantic-bokmål: *Kompisen hans i Milano var på Tinder-dates og drev på med forskjellig.*
WER: 50%
SemDist: 0.03

(5)

lydfil: *Tro_Den_andre_siden_250113_906000_1082000_s37425_e44914.wav*
gullstandard: ***Og da var*** *min bror og pappa* ***der inne*** *hos mamma,* ***og så ringte*** *de etter en time og sa at* ***nå må du komme.***
nb-whisper-large-semantic-bokmål: *Min bror og pappa var hos mamma. De ringte etter en time og sa at* ***jeg måtte komme.***
WER: 44%
SemDist: 0.07

I (4) har modellen omskrevet starten av setninga slik at det ikke er subjektsdublering (*han kompisen…han*) eller pseudokoordinering (*dreiv jo og var*), to konstruksjoner som er mer vanlig muntlig enn skriftlig. Ingen av delene påvirker betydninga direkte, så man ender opp med en transkripsjon som betyr det samme som det som sies i lydfila. I (5) er omskrivinga noe mer omfattende. Transkripsjonen gjengir mye av meningsinnholdet, men et stedsadverbial, *der inne*, og et tidsadverbial, *nå*, er utelatt, som gjør transkripsjonen noe mindre spesifikk. Merk også at utfyllinga til utsagnsverbet *sa* er endra fra en blanding av direkte og indirekte tale (*mixed quotation*) *at nå må du komme* til normal indirekte tale *at jeg måtte komme*. Begge konstruksjoner er korrekte i denne konteksten, og med unntak av det utelatte tidsadverbialet har endringa blitt riktig. Denne endringa vitner om at modellen har evner til å omskrive ganske betydelig fra det som faktisk sies uten at betydninga endres substansielt.

Andre eksempler fra nb-whisper-large-semantic-bokmål er mer problematiske. I (6) er transkripsjonen på nynorsk, selv om modellen er satt til å transkribere utelukkende på bokmål. Dette fanges trolig ikke opp av SemDist, fordi modellen som ligger til grunn for utregninga av SemDist både er trent på bokmåls- og nynorskdata. I (7) har subjektet blitt

utelatt to ganger, noe som gjør transkripsjonen ugrammatisk. Det er også annen betydningsfull informasjon som er utelatt, som for eksempel fornavnene til trenerne og vurderinga av dem som fantastiske.

(6)

lydfil:
Sammen_med_Sandvik_Vibeke_K_Ottesen_030121_137000_515000_s305540_e323150.wav
gullstandard: *Og hvis da det at jeg var i den situasjonen jeg var i blei et signal for henne, så er jo hun, som del av menneskeheten, også da en del av en art som er evolvert til å trekke seg unna.*
nb-whisper-large-semantic-bokmål: *Viss situasjonen eg var i, blei eit signal for henne, er ho som del av menneskeheita også ein art som er evolvert til å trekke seg unna.*
WER: 48,78%
SemDist: 0.04

(7)

lydfil:
Timeout_Episode_19_Preben_Vildalen_100619_143000_558000_s156410_e175600.wav
gullstandard: *Ja **da, så** har det vært perioder med landslag**et hvor det har vært ymse med vært** mye rot, men jeg er **veldig** stolt over det, **og jeg** har 200 landskamper og**, som du sier,** hatt **fra Harald** Madsen **til Gunnar** Pettersen som trenere, **som to fantastiske trenere det og ja, stolt av det.***
nb-whisper-large-semantic-bokmål: *Ja, det har vært perioder med landslag med mye rot, men jeg er stolt over det. Har 200 landskamper og hatt Madsen og Pettersen som trenere.*
WER: 56,60%
SemDist: 0.04

I de fire eksemplene gjengitt her er WER på rundt 50%, som vil si at annethvert ord er galt. Det er å regne som svært dårlig WER. Selv om det er problemer med noen av disse transkripsjonene, gjengis likevel meningen forholdsvis godt, med unntak av i det siste eksempelet. På grunn av at de automatiske transkripsjonene har en mer skriftlig stil, kan man argumentere for at de er mer nyttige for en del applikasjoner, for eksempel TV-teksting, enn helt ordrette transkripsjoner ville vært. For eksemplene (4) til (6) ser SemDist ut til å være et mer nyttig mål enn WER. (7) viser imidlertid at heller ikke SemDist er et fullgodt mål for å identifisere dårlige transkripsjoner.

For utvalget fra usm er forventninga at vi skal se transkripsjoner der ikke så mange ord er endra, men der endringene påvirker betydninga substansielt. Her har vi nemlig plukka eksempler med høy SemDist og relativt lav WER. Denne forventninga stemmer godt med de faktiske observasjonene: 9 av 10 eksempler er vanskelige å forstå, selv om det er relativt få endringer av ord. I (8) har det formelle subjektet *det* blitt endra til det personlige pronomenet *du*, og stedsnavnet *Kautokeino* har ikke blitt gjenkjent. I (9) har to viktige ord i setninga, *letta* og *skade*, blitt feilgjenkjent, og resultatet er en ugrammatisk transkripsjon.

(8)

lydfil: *Drivkraft_Elle_Marja_Eira_130922_28000_898000_s167079_e169359.wav*

gullstandard: *Ja, og da spiller **det** jo ikke noen rolle at man bor i **Kautokeino**?*
usm: *ja og da spiller **du** ikke noen rolle at man bor i **kautskyen***
WER: 21,43%
SemDist: 0.37

(9)

lydfil: Distriktsnyheter_Rogaland_290722_0_539000_s139450_e147740.wav
gullstandard: *Når hendelsen skjedde, så var jo vi mest av alt **letta** fordi at vi ikke hadde **skade** på hverken liv eller helse.*
usm: når hendelsen skjedde så var jo vi mest av alt **lette** fordi at vi ikke hadde **skate** på verken liv eller helse
WER: 13.64%
SemDist: 0.28

Begge disse eksemplene har forholdsvis høy WER, men likevel langt lavere enn eksemplene fra nb-whisper-large-semantic-bokmål vi gjenga over. Imidlertid er transkripsjonene ikke brukbare slik de er. De relativt høye SemDist-verdiene røper imidlertid at det er noe galt med disse transkripsjonene.

Vi har sett at ikke-ordrette, men meningsfulle transkripsjoner kan få høye WER-verdier, og at transkripsjoner med lav WER kan ha alvorlige feil som påvirker betydningen. SemDist ser i noen grad ut til å fange opp slike tilfeller. Vi tror likevel ikke at SemDist alene er et godt mål på kvaliteten til talegjenkjenningssystemer. For det første har ikke SemDist en transparent fortolkning, som tidligere nevnt, fordi utregningen baserer seg på en maskinlæringsmodell. For det andre er det tilfeller som ikke er brukbare, som (7), som får relativt lav SemDist på tross av at transkripsjonen ikke er brukbar. Siden SemDist ikke har en transparent fortolkning, er det vanskelig å si noe presist om hvorfor dette skjer eller å kontrollere for dette. For det tredje er det ikke nødvendigvis tilstrekkelig at en transkripsjon betyr det samme som det som sies. Mange brukere vil nok også ha en forventning om at transkripsjonen i hovedsak bruker de samme ordene som talen, selv om dette kan variere fra brukstilfelle til brukstilfelle. Eksperimentet med å hente ut 10 setninger som enten har lav SemDist og høy WER eller høy WER og lav SemDist, viser at det kan ha en verdi å se disse målene sammen. Ikke-ordrette systemer vil ha noe høyere gjennomsnittlig WER enn ordrette systemer av samme kvalitet, men de bør også ha relativt lav SemDist. Ordrette systemer bør ha lav WER, men hvis de i tillegg har høy SemDist, kan det være en indikasjon på at det er en del betydningsendrende feil i transkripsjonene.

Det viktigste kvalitetsmålet for et talegjenkjenningssystem er til syvende og sist hvor godt det fungerer for brukerne av det, og svaret på det får man ved å spørre brukere. Et ideelt automatisk mål for talegjenkjenning bør ligge så nært som mulig en slik menneskelig vurdering. Openai-whisper-large-bokmål, et system som har fått mange brukere og som mange er tilfredse med, får i vår undersøkelse en høy gjennomsnittlig WER (30,7%) og har en gjennomsnittlig SemDist som ligger midt på treet sammenlikna med de andre systemene vi har testa (0,20). Våre automatiske mål ser med andre ord ut til å samsvare dårlig med den opplevde kvaliteten på dette systemet. Vi har inspisert 10 segmenter med relativt høy WER og lav SemDist for openai-whisper-large-bokmål på samme måte som vi gjorde for nb-whisper-large-semantic-bokmål. 6 av 10 segmenter er korte segmenter, som får høy WER fordi de inneholder få ord. Det er ikke noe klart mønster å se i de resterende

segmentene. Vi vil se seinere i rapporten at denne modellen *hallusinerer* en del, det vil si produserer transkripsjoner som ikke gjengir det som sies overhodet. Det er sannsynlig at disse hallusineringene er en del av grunnen til at openai-whisper-large-bokmål kommer såpass dårlig ut i våre tester.

Som konklusjon vil vi si at WER er et mer meningsfullt mål for modeller som transkriberer ordrett enn for modeller som ikke gjør det. WER gir et mål på hvor nøyaktige ordrette modeller er, som er en nyttig indikasjon på kvalitet. Enkelte ordfeil kan imidlertid ha stor betydning på forståelsen av en setning, og transkripsjoner kan derfor være vanskelig å forstå på tross av lav WER. SemDist kan derfor være et nyttig kvalitetsmål i tillegg til WER også for ordrette modeller. For ikke-ordrette modeller vil WER være høy sammenlikna med ordrette modeller med tilsvarende kvalitet. WER gir derfor begrensa informasjon om kvaliteten til disse modellene. Likevel er det antakelig et kvalitetstegn i de fleste tilfeller at transkripsjoner ligger nært opp til hva som ble sagt, så WER har sin plass i evalueringa slike systemer også. Det er også et transparent og forklartbart mål i motsetning til SemDist. SemDist kan gi nyttig supplerende informasjon for ikke-ordrette modeller, men man har ingen garanti for at transkripsjoner med lav SemDist er korrekte. Det er altså vanskelig å tallfeste på en god måte kvaliteten til ikke-ordrette systemer med de kvalitetsmålene vi har til rådighet. I det følgende vil vi likevel bruke WER og SemDist som kvalitetsmål, fordi det er det beste vi har og fordi de ser ut til å gi nyttig informasjon også om ikke-ordrette modeller.

Hvilke feil gjør norske talegjenkjenningssystemer

I dette underkapittelet skal vi se nærmere på hvilke faktorer som påvirker kvaliteten på transkripsjonene fra de forskjellige modellene. Vi gjør dette for å gi et bedre bilde av hvordan systemene fungerer, men også for å peke ut hvor det er forbedringspotensial og hvor Språkbanken og andre relevante aktører kan bidra.

2 a) Hvor gode er systemene til å håndtere ulike dialekter, kjønn, opptaksforhold og overlappende tale?

Problemstilling 2 a omhandler predefinerte faktorer som kan ha en effekt på kvaliteten på talegjenkjenning: dialekt, kjønn, opptaksforhold og overlappende tale.[13]

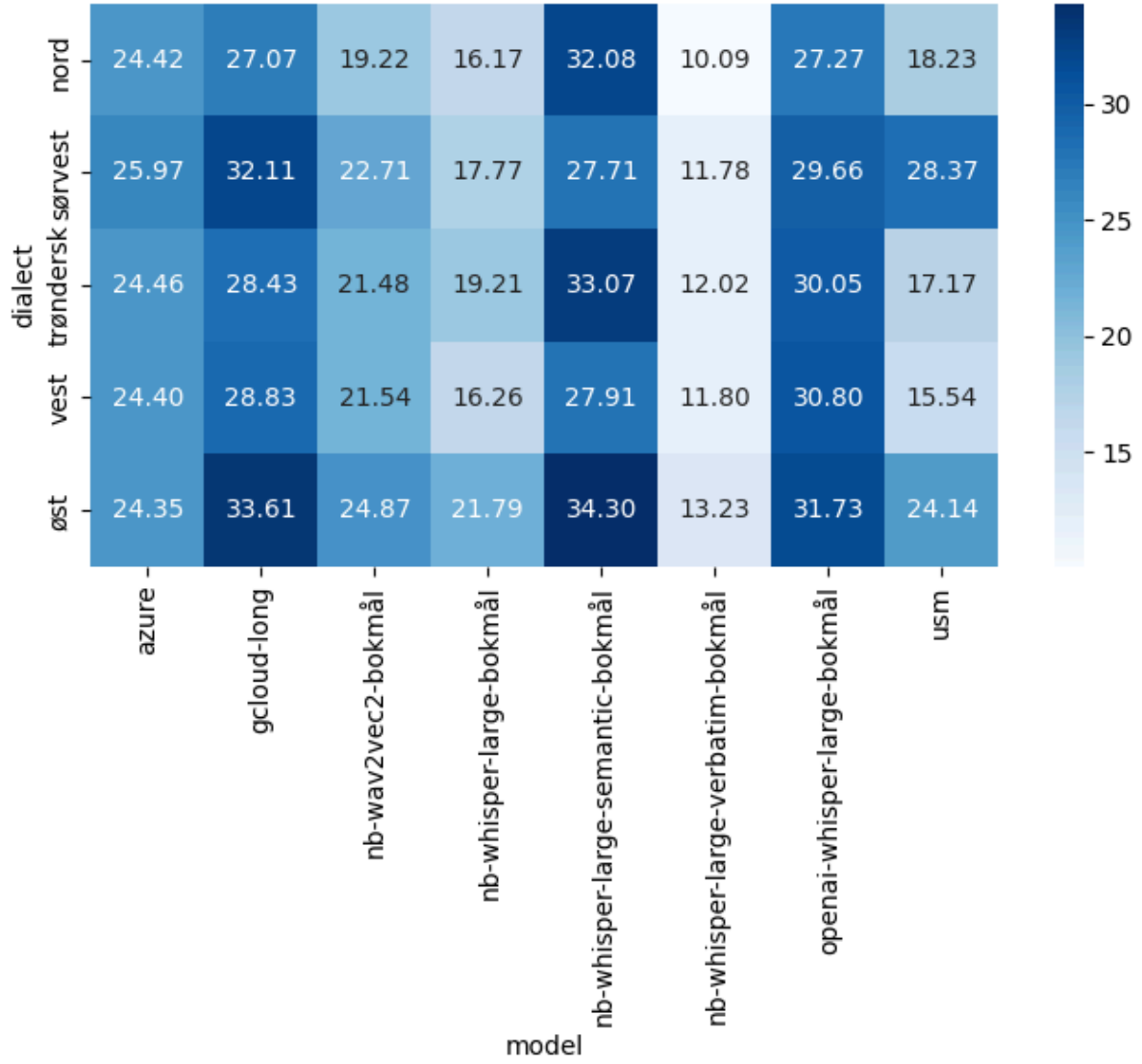
Dialekt

Som forklart over er hver taler i testsettet annotert med en dialektregion. De er også annotert med finkorna dialektområder. Vi har brukt denne informasjonen til å måle gjennomsnittlig WER per dialektregion og per dialektområde. Vi har valgt å skille bokmål- og nynorsk i denne undersøkelsen, da vi mistenker at nynorskmodellene er trent på mer data fra nynorskområder og favoriserer dialekter derfra.

[13] *Semantisk kompleksitet* var egentlig også på denne lista, men vi fjerna det, fordi vi ikke fant et brukbart mål på dette. Før vi valgte å ta det ut, testa vi stoppordsratio og perpleksitet, og ingen av de målene korrelerte betydelig med WER eller SemDist.

Bokmål

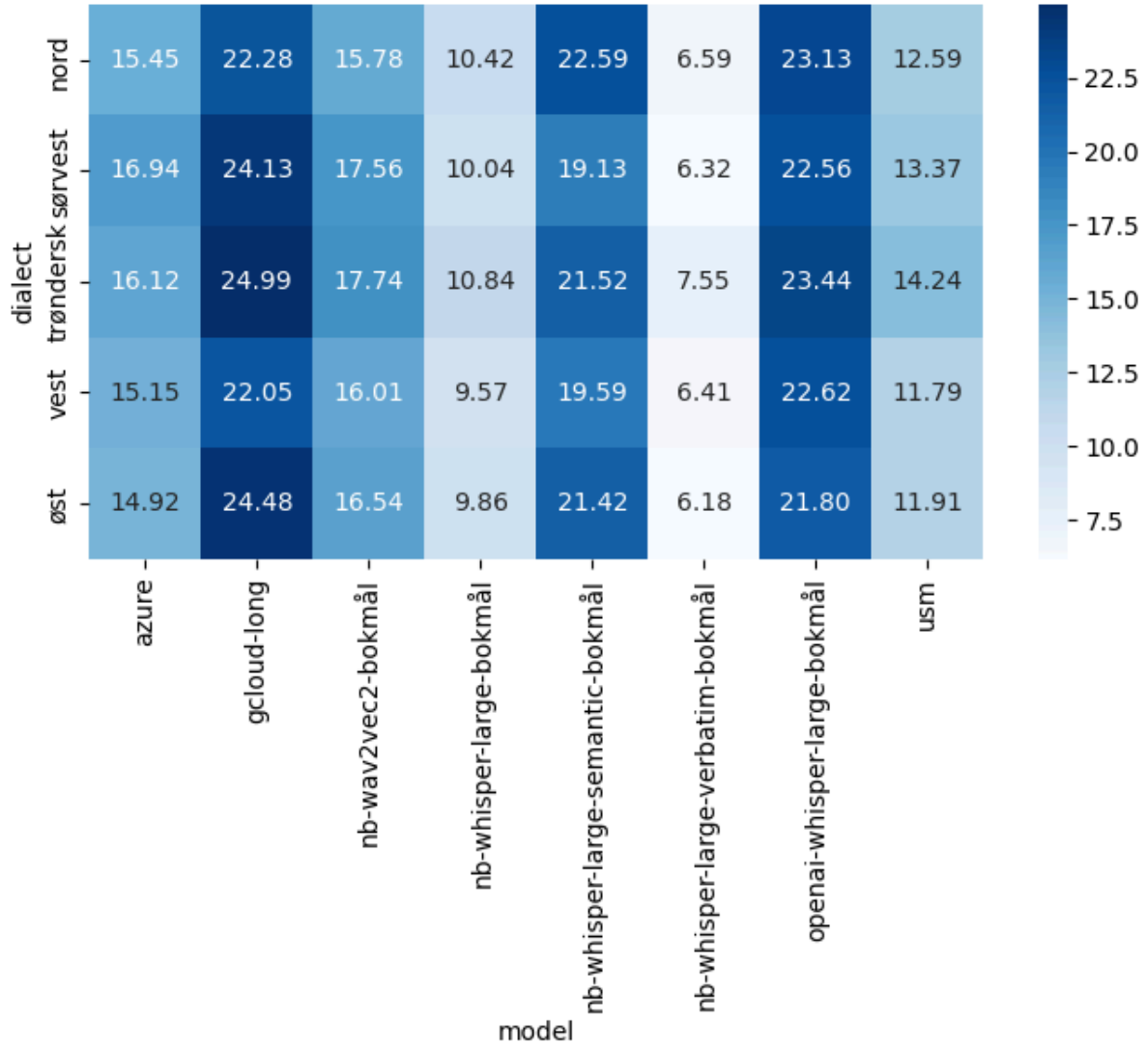
Figur 1 viser gjennomsnittlig WER for hver dialektregion på hele datasettet for bokmålsmodellene. Figuren er fargekoda etter hvor høy ordfeilraten er.



Figur 1: WER for hver dialektregion

Usm har størst differanse mellom høyeste og laveste WER (12,83) med score for sørvestnorsk og østnorsk som ligger betydelig over de andre områdene. Azure har relativt liten differanse mellom høyeste og laveste WER (1,62), og utslagene er også forholdsvis lave for nb-whisper-large-verbatim (3,14). De andre dialektene har en differanse mellom 6,59 og 4,46. Østnorsk har relativt sett høy gjennomsnittlig WER for alle modellene utenom Azure. Dette er et overraskende resultat, fordi de fleste modellene er trent på mye østnorsk og fordi mange varianter av østnorsk har en uttale som ligger nært opp til bokmål. Dette resultatet stemmer også dårlig overens med resultater vi har sett på andre datasett, der trøndersk og vestlandsdialekter pleier å være de vanskeligste dialektene. Vi mistenker at de relativt høye resultatet for østnorsk skyldes skjevheter i datautvalget heller enn dialekt. Østnorsk har flere unike talere enn de andre dialektene (119, mens de andre har mellom 46

og 80), og har også mest taletid (29%, mens de andre dialektene ligger mellom 15% og 19%). Det er derfor sannsynlig at de østnorske dataene har større kompleksitet enn dataene for de andre dialektene, og at dette drar WER opp. For å undersøke dette nærmere, har vi sett på gjennomsnittlig WER for ulike datautvalg der segmentene med høyest WER er holdt ute. Figur 2 viser gjennomsnittlig WER for de 75% beste segmentene for hver modell.

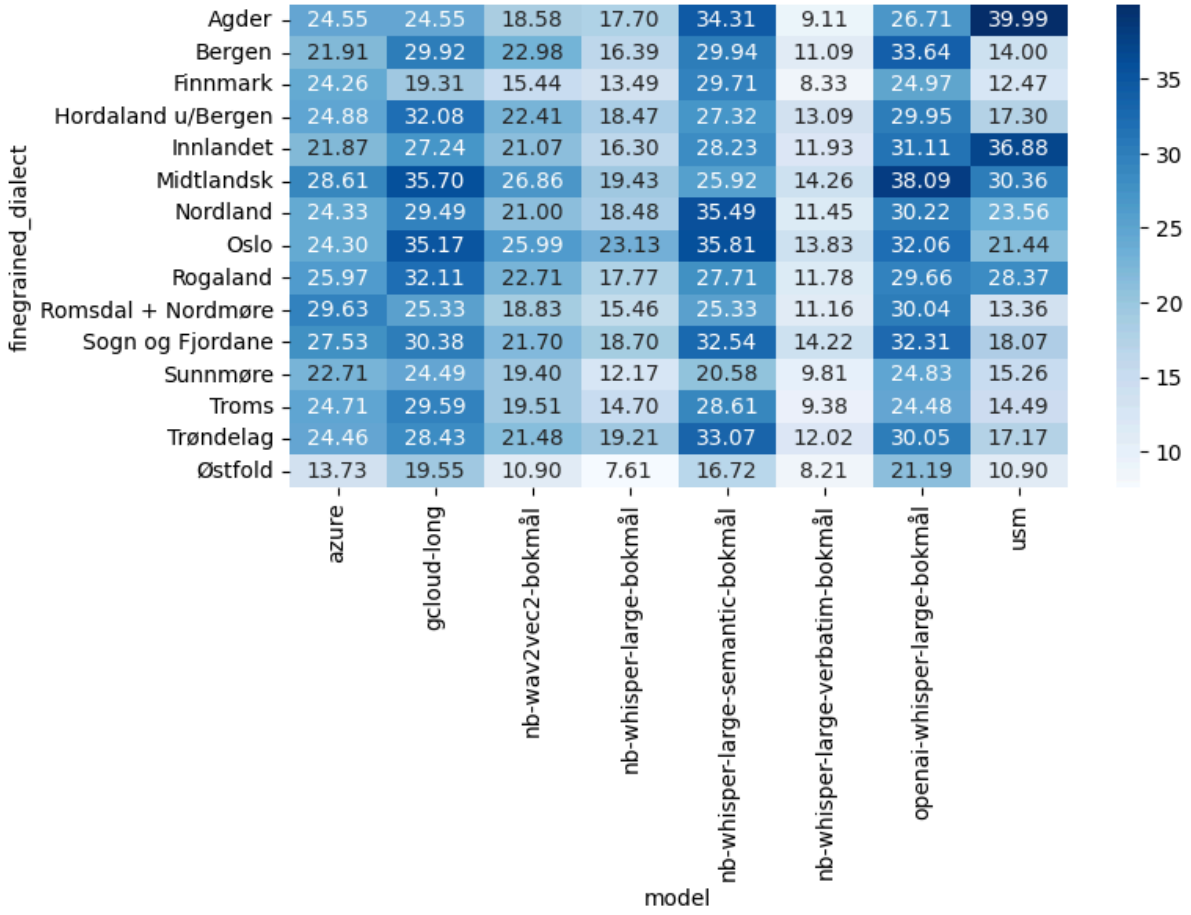


Figur 2: Gjennomsnittlig WER for de 75% beste segmentene for hver modell

Her ser vi for det første at WER verdiene er mindre spredt. For det andre forsvinner de relativt høye WER-verdiene for østnorsk for alle modellene når vi fjerner dataene med høyest WER. Det kan tyde på at problematiske segmenter er overrepresentert for østnorsk heller enn at modellene sliter med østnorsk. Trøndersk har tatt over ledelsen som dialekta med høyest WER for de fleste modellene i dette datautvalget. Dette gir mening, da trøndersk avviker ganske mye fra bokmålsnormen i morfologi og uttale.

Alt i alt ser dialekt ut til å ha en viss effekt på WER, men utslagene er ikke veldig store når man holder til side dataene med høyest WER.

Vi har også sett på WER for de mer finkorna dialektkategoriene i testsettet. Figur 3 gir gjennomsnittlig ordfeilrate på hele datasettet gruppert på disse kategoriene.



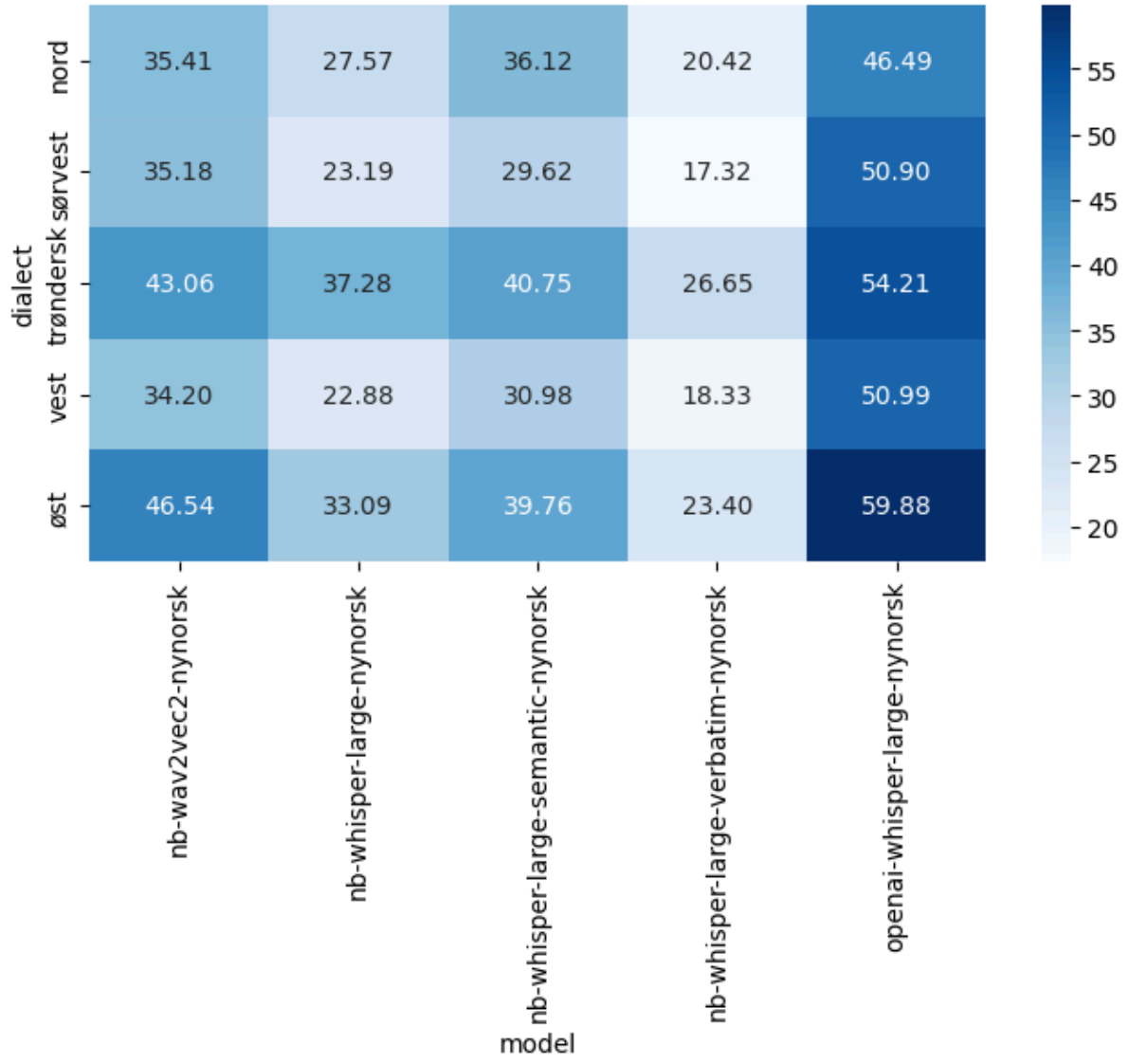
Figur 3: Gjennomsnittlig WER for finkorna dialektområder

Her ser vi mye større differanser mellom beste og dårligste dialekt for flere systemer. Dette gjelder spesielt for usm med en differanse på 29,09. nb-whisper-large-verbatim-bokmål har en forholdsvis lav differanse sammenlikna med de andre systemene (6,05). De andre modellene har en differanse mellom 15,52 og 19,09. Merk imidlertid at en del av disse kategoriene inneholder ganske få talere. Dermed er det sannsynlig at særegenheter ved enkelttalere og opptakene av disse kan ha stort utslag i positiv eller negativ retning på resultatene. Det er derfor trolig ikke mulig å si med utgangspunkt i disse resultatene hvilke enkeltdialekter modellene er spesielt gode eller dårlige på.

En interessant observasjon er at differansen mellom dialekter, både grovkornede og finkornede, er relativt liten for nb-whisper-large-verbatim-bokmål. Dette er en modell som transkriberer ordrett, og WER fungerer bedre som mål for kvalitet på den enn på de andre NB Whisper-modellene. Når denne modellen får såpass like resultater, indikerer det at Whisper-teknologien og dataene NB Whisper er trent på, kan gi modeller som håndterer norsk dialektvariasjon på en god måte.

Nynorsk

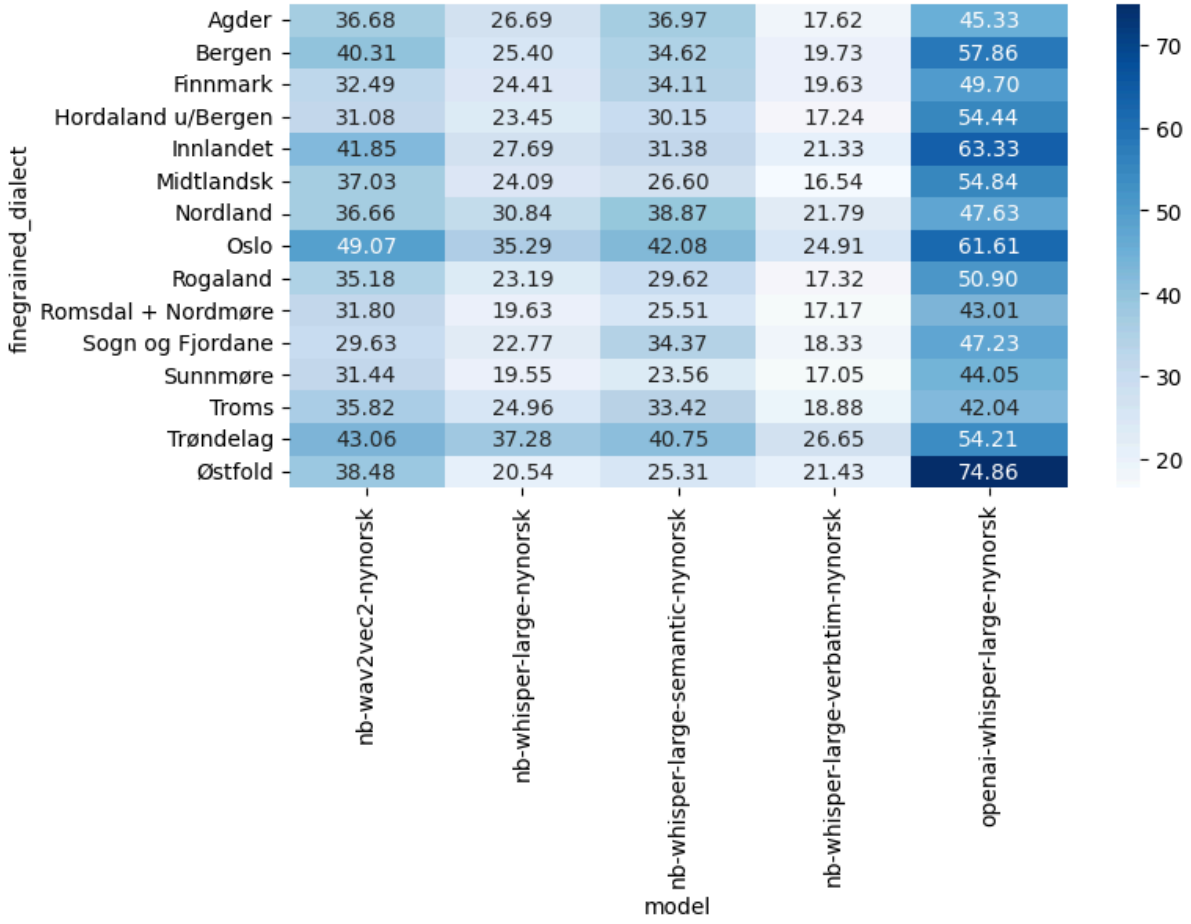
I figur 4 ser vi variasjonen i WER mellom dialekter på nynorsk på hele datasettet.



Figur 4: Gjennomsnittlig WER per dialektregion - nynorsk

På tvers av systemene er vestnorsk og sørvestnorsk dialektene med lavest WER. Mens østnorsk, trøndersk, og til dels nordnorsk, har en del høyere resultater. Dette er ikke overraskende. Kjerneområdene for nynorsk er på vestlandet, så alle systemene har trolig blitt trent på mest nynorskdata fra disse områdene. I tillegg er mange dialekter på vestlandet og sørvestlandet relativt nynorsknære, mens for eksempel oslomål er ganske fjernt fra nynorsk. Vi ser at forskjellen mellom dialektområdene med lavest og høyest WER er relativt stor for alle systemer, fra 14,40 for nb-whisper-large-nynorsk til 9,33 for nb-whisper-large-verbatim-nynorsk. Vi har også sjekka fordelinga for de 75% beste segmentene for hvert system på samme måte som for bokmål. Vi ser de samme tendensene i dette utvalget som i de totale dataene.

Figur 5 viser gjennomsnittlig WER på nynorsk for finkorna dialektområder.



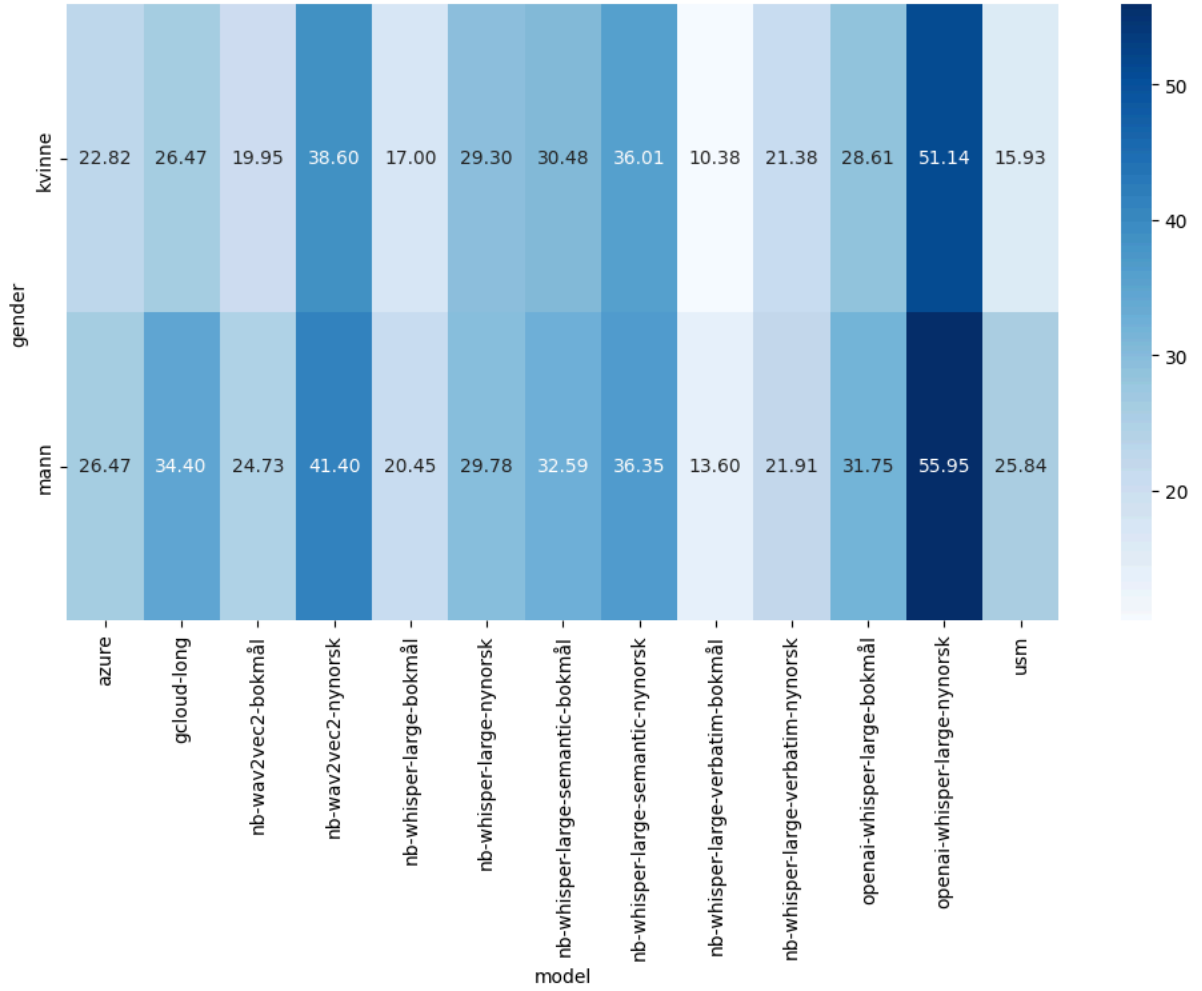
Figur 5: Gjennomsnittlig WER for finkorna dialektområder - nynorsk

Som nevnt over blir utvalget av talere såpass lite for hver dialekt at det er vanskelig å dra bastante konklusjoner, men vi ser at nynorsknære dialekter som midtlandsk og vestlandsdialekter utenom Bergen har relativt mye bedre resultater enn Oslo og Trøndelag.

For nynorsk ser vi altså at dialekt påvirker resultatene i større grad enn for bokmål. Dette er ikke overraskende, gitt tilgangen på treningsdata. Man bør likevel forvente av gode talegjenkjenningssystemer for nynorsk at de transkriberer alle dialekter relativt korrekt. På dette området ser det derfor ut til å være et klart forbedringspotensial.

Kjønn

Figur 6 viser gjennomsnittlig WER for menn og kvinner for de ulike modellene.

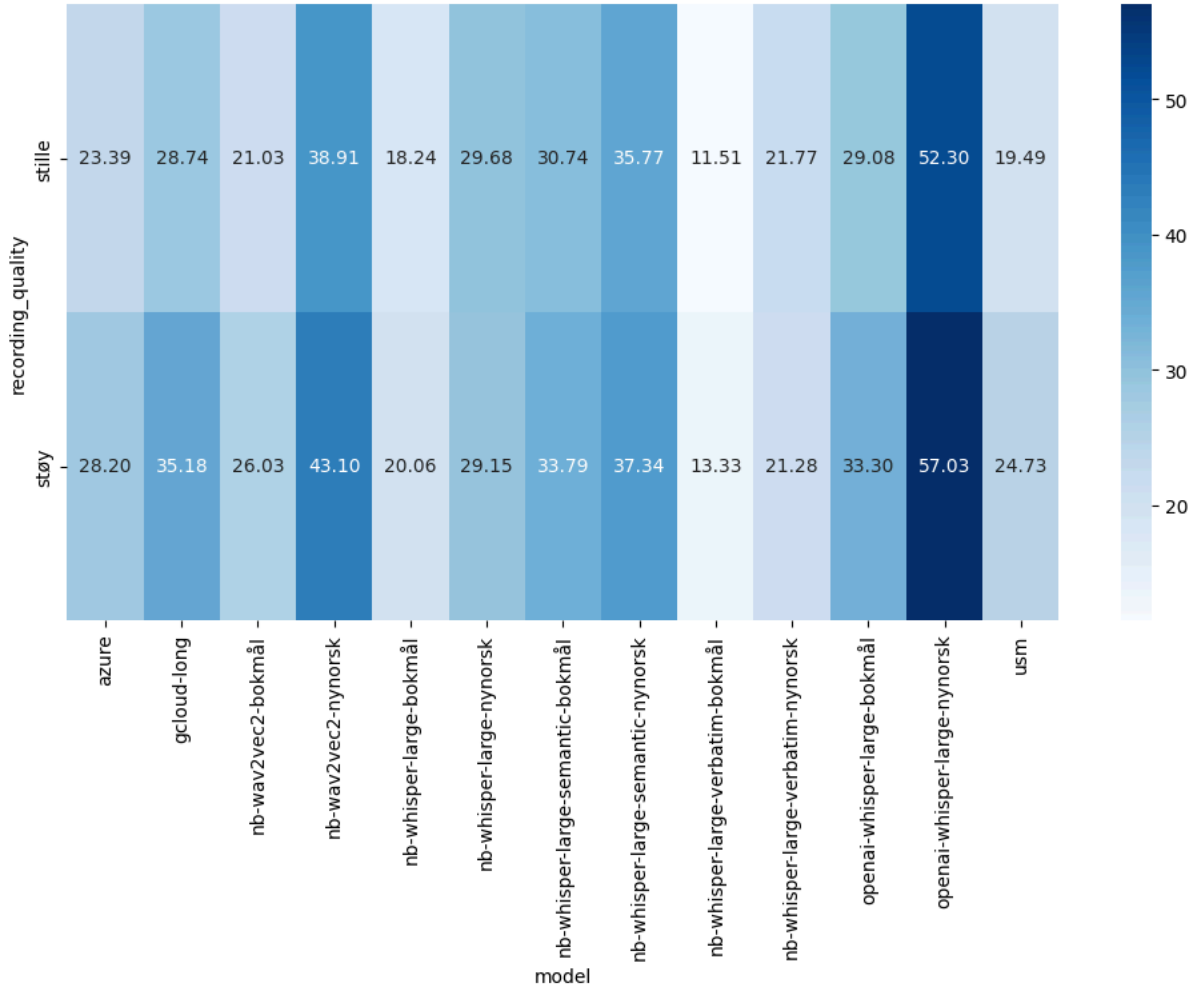


Figur 6: gjennomsnittlig WER fordelt på kjønn

Vi ser at ordfeilraten gjennomgående er lavest for kvinner for bokmålssystemene. For usm og gcloud-long er denne forskjellen på 9,91 og 7,73, mens de andre bokmålssystemene har en forskjell mellom 2,11 og 4,78. Vi ser også en tilsvarende forskjell for wav2vec2 og OpenAI Whisper på nynorsk, mens forskjellen mellom kjønn er relativt liten for NB Whisper på nynorsk. At mange av systemene gjør det bedre for kvinner enn for menn er overraskende, gitt at de fleste treningssett for talegjenkjenning har noe flere menn enn kvinner. Vi vet ikke sikkert hva denne forskjellen kan komme av. Som for dialekt er det mulig at en del av forklaringen kan v’re at de to kategoriene ikke er helt likt fordelt i testsettet. Det er 199 menn i testsettet og 161 kvinner. Taletida er imidlertid bortimot lik (50,7% kvinner og 49,3% menn). Vi ser også samme fordeling når vi ser på de beste 75% av dataene for hver modell, selv om utslagene er mindre dramatiske. Interessant nok har en tilsvarende ubalanse til fordel for kvinner har blitt observert for andre språk også (se Feng et al, 2024 og referanser der). Dette kan være et interessant emne for videre forskning, men siden forskjellen mellom kjønnene er relativt liten for flere av systemene, tror vi ikke at man trenger å prioritere å utjevne denne forskjellen fullstendig.

Opptaksforhold

Alle segmenter er annotert med informasjon om opptakskvalitet. Transkribørene har hatt tre kategorier: *støy*, *stille* og *telefon*, hvor *støy* indikerer bakgrunnsstøy og *telefon* indikerer telefonkvalitet. Det var imidlertid så få opptak med telefonkvalitet at vi valgte å slå sammen kategoriene *stille* og *telefon* i denne undersøkelsen. Figur 7 viser den gjennomsnittlige ordfeilraten for *stille* og *støy*:



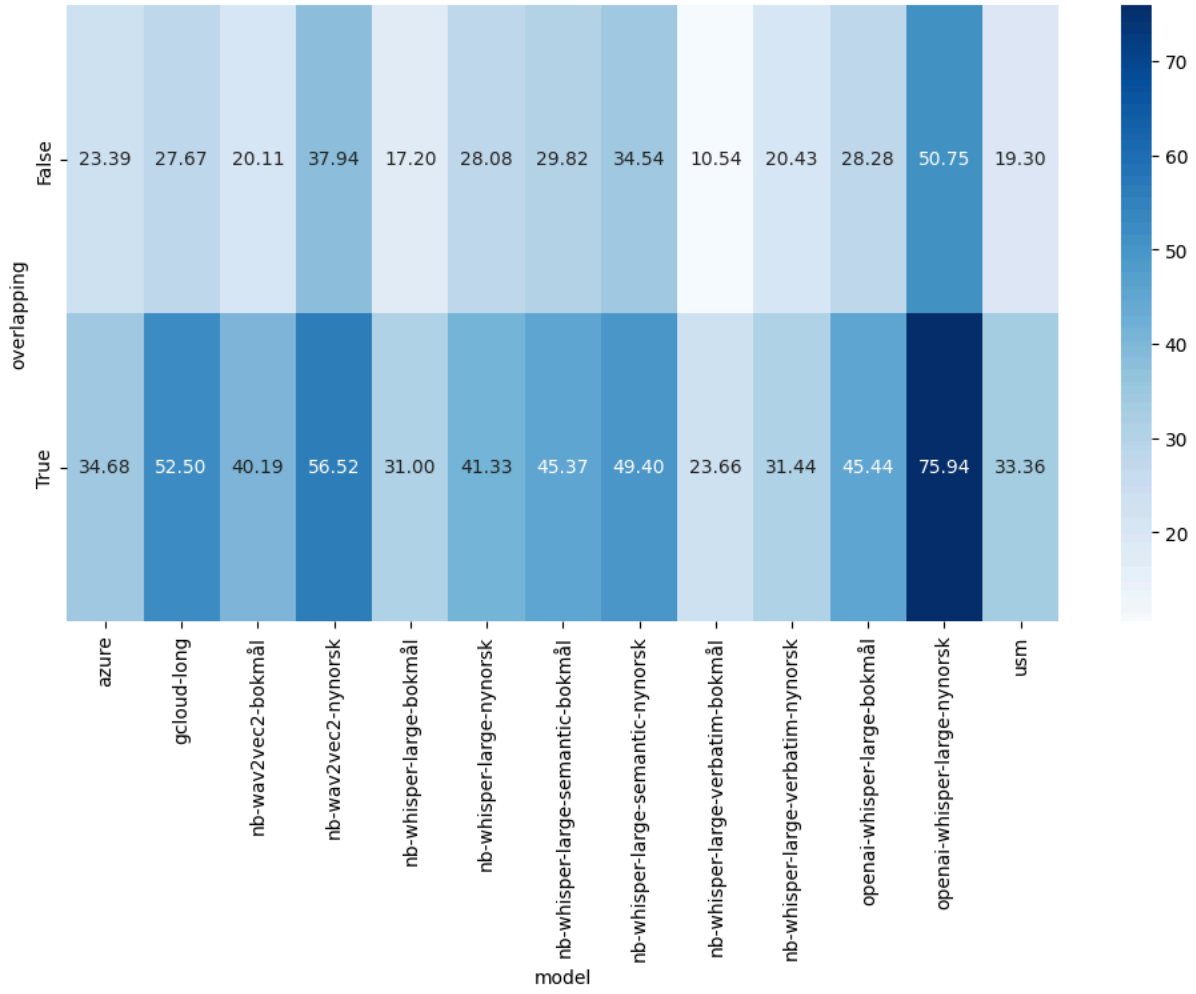
Figur 7: Gjennomsnittlig WER fordelt på opptaksforhold

Bakgrunnsstøy er en kjent faktor som påvirker talegjenkjenning negativt, og ikke overraskende har de fleste systemene høyere WER i opptak med støy. Med unntak av nb-whisper-large-semantic-bokmål er det imidlertid relativt små forskjeller for NB Whisper-systemene, og nb-whisper-large-nynorsk og nb-whisper-large-verbatim-nynorsk har faktisk marginalt bedre resultater for opptak med støy enn opptak uten støy.

Overlappende tale

Dersom talegjenkjenningssystemer skal kunne håndtere dialog på en god måte, bør de kunne transkribere overlappende tale ganske nøyaktig. Dette er imidlertid en vanskelig oppgave både for maskiner og mennesker, blant annet fordi det er vanskelig å skille ut hvilken tale som er relevant. Vi ville teste i hvilken grad overlappende tale påvirker WER.

Dersom starten på et segment er før slutten på det forrige, eller slutten på et segment er etter begynnelsen på det neste, merker vi det som overlappende.[14] 14,23% av segmentene har overlappende tale. Gjennomsnittlig WER for segmenter med og uten overlappende tale vises i figur 8:



Figur 8: Gjennomsnittlig WER med og uten overlappende tale

Ikke overraskende fører overlappende tale til høyere WER for alle modeller. Effekten er kraftigere enn de andre faktorene vi har sett på. Differansen i WER mellom overlappende og ikke-overlappende tale går fra 25,19 for openai-whisper-large-nynorsk til 11,01 for nb-whisper-large-verbatim-nynorsk.

Merk at det i en del tilfeller med overlappende tale ikke er åpenbart for et menneske hvilken tale som skal transkriberes. I slike tilfeller har vi observert at systemene noen ganger har transkribert korrekt en annens tale enn den som er transkribert i gullstandarden. Hvor ofte dette forekommer, vet vi imidlertid ikke.

[14] Vi runda av start- og slutt-tidskoder til et heltall, så overlappet må være på minst et halvt sekund for at det skal telle.

2 b) Hvilke andre faktorer påvirker kvaliteten på transkripsjonene til de forskjellige systemene?

I delkapitlene over tok vi utgangspunkt i faktorer vi mistenker at påvirker kvaliteten på talegjenkjenning og sjekka i hvilken grad de gjorde det. I dette delkapittelet inspiserer vi transkripsjonene på forskjellige måter og prøver å identifisere andre faktorer som påvirker kvaliteten. Vi vil bruke tre metoder for å gjøre dette:

1. **Nøkkelordsanalyse**: Vi vil bruke en metode fra korpuslingvistikken som kalles *nøkkelordsanalyse* (*keyword analysis*). I sin vanlige bruk henter man ut frekvensordlister fra en tekst man ønsker å analysere og sammenlikner disse med frekvensordlista til et referansekorpus. På den måten kan man generere en liste med ord som enten har en overraskende høy frekvens i teksten gitt referansekorpuset eller en overraskende lav frekvens (Pojanapunya & Watson, 2018). Metoden kan imidlertid også brukes for å oppdage særegenheter ved talegjenkjenningssystemer. Da bruker man gullstandarden som referansekorpus og ser på ord med en overraskende høy eller overraskende lav frekvens i transkripsjonene til en modell (Solberg et al., 2023).
2. **Stikkprøver:** Vi studerer tilfeldig utvalgte segmenter fra hele datasettet for å finne typiske mønster.
3. **Feilanalyse:** Det kan også være nyttig å vite hva som skjer når systemene gjør det virkelig dårlig. Vi vil derfor se på stikkprøver fra de 500 segmentene med høyest WER.

azure

Modellen transkriberer ord for ord med store og små bokstaver og tegnsetting. Det ser ikke ut til å være noen forenklende omformuleringer av den typen vi ser fra Whisper-modeller. Når modellen ikke klarer å transkribere riktig, skriver den gjerne ord som er korrekt stavede ord på norsk, men som ikke samsvarer med det som sies. Slike gale ord passer ofte ikke i konteksten, betydningsmessig eller grammatisk, slik at man får meningsløse ordsekvenser. Modellen produserer også en del særskriving, som *boplikt fritak* istedenfor *bopliktfritak*. Vi ser også at tall ofte ser ut til å bli skrevet med siffer, selv i uttrykket *tusen takk*, som blir transkribert som *1000 takk*. Tallord mellom 1 og 12 blir gjerne blir skrevet med bokstaver i korrekt, skriftlig norsk. Vi ser imidlertid at azure har en tendens til å skrive også disse med siffer. Selv om modellen hadde relativt lik WER for ulike dialektområder, ser vi i stikkprøvene en del tilfeller der modellen misforstår dialektuttale.

For en del av segmentene våre har azure slutta å transkribere for tidlig, så siste del av talen ikke er med. Vi har imidlertid sendt modellen ett og ett segment istedenfor å gi den én stor lydfil, som man vil gjøre i de fleste applikasjoner. Vi vet ikke om dette problemet også oppstår når man gir modellen store lydfiler eller om det har å gjøre med at vi sender den korte lydfiler. Feilanalysen viser at en del veldig korte segmenter ikke har noen transkripsjon overhodet. Dette kan hende at dette problemet også er knytta til at vi transkriberer én lydfil per segment.

gcloud-long

Transkripsjonene fra denne modellen er ordrette med store og små bokstaver, men ikke tegnsetting. Bruken av store og små bokstaver følger imidlertid ofte ikke norsk ortografi og virker litt tilfeldig satt. Merk for eksempel *by Gård*, *Ok* og *Hjelp* i (10).

(10)

lydfil: 72_timer_Silje_Nordnes_181018_0_881000_s704610_e719891.wav
gullstandard: *Men altså i fjerde etasje i en bygård i Oslo, har du mye kontakt med naboen, har du tenkt på at, ok, han eller hun trenger kanskje noe ekstrahjelp hvis noe skulle skje, eller hun eller han kan jeg spørre om, kan jeg få låne ei dyne, eller hvordan er samholdet?*
gcloud-long: *men altså i fjerde etasje i en by Gård i Oslo Har du mye kontakt med naboen har du tenkt på at Ok han eller hun trenger kanskje noe ekstra Hjelp hvis dere skal se eller hun eller han kan jeg spørre om hva kan jeg få låne dine eller hvordan er det samhold*
WER: 23,53%
SemDist: 0,06

Det er ellers noe særskriving, som *by Gård* i (10). Som azure bruker gloud-long gjerne korrekt stavede ord når den feiler, noe som fører til meningsløse ordsekvenser i en del tilfeller. Modellen har relativt høy gjennomsnitttlig ordfeilrate (30,40%). Med utgangspunkt i stikkprøvene mistenker vi at én faktor påvirker dette snittet i ganske stor grad: Transkripsjonene mangler svært ofte for begynnelsen og slutten av segmentene, og korte segmenter får ofte ikke en transkripsjon i det hele tatt. Som nevnt over kan dette være forårsaket av at vi transkriberer små lydfiler av gangen, og det er mulig man ikke hadde hatt de samme problemene om man hadde transkribert større lydfiler. I flere av tilfellene vi så på, var det meste korrekt utenom at begynnelsen og/eller slutten mangla.

usm

Usm produserer en ordrett transkripsjon med kun små bokstaver og ingen tegnsetting. I nøkkelordsanalysen for usm ser vi at denne modellen av og til produserer utenlandske ord som *nein*, *nu* og *nej*. Transkripsjonene ser ofte ut til å være veldig nøyaktige, og inkluderer en del transkripsjon av nølelyder og gjentatte ord. Usm kan av og til feilstave ord, særlig ved ikke-bokmålsnære dialekter. Tall blir ofte transkribert med siffer, selv når det er snakk om tall under 12.

I feilanalysen ser vi at en del segmenter ikke har blitt transkribert i det hele tatt. Det er imidlertid ikke åpenbart at lengden på segmenter er en faktor, i motsetning til azure og gcloud-long. Begynnelsen og slutten på segmenter ser også ut til å være med. Vi ser at noen gale transkripsjoner har svenske ord.

Alt i alt framstår usm som mer nøyaktig enn de to andre kommersielle modellene, azure og gcoud-long. Den ordrette transkripsjonen uten tegnsetting og med nølelyder og gjentakende ord kan være nyttig for en del bruksområder der nøyaktig referat er viktig. Det er kan hende mindre nyttig for for eksempel undertekst.

nb-wav2vec2-bokmål

Nøkkelordsanalysen viser at transkripsjonene til wav2vec2-modellen inneholder mange enkeltbokstaver omgitt av mellomrom som ikke står i gullstandarden. Det er dessuten en del tilfeller av nynorskordene *kjem* og *eg*.

Stikkprøvene likner en del på dem fra usm: Transkripsjonene er ordrette med kun små bokstaver og ingen tegnsetting. Gjentatte ord blir transkribert. Transkripsjonene er ofte nesten korrekte, men modellen produserer en del skrivefeil. Dette ser særlig ut til å forekomme der dialekten avviker fra skriftnormen, og skrivefeilene ligger tett opp til uttalen. Tall transkriberes konsekvent med bokstaver.

Feilanalysen viser at en del opptak blir transkribert med meningsløse strenger med bokstaver eller én enkelt bokstav.

Dette kan være en nyttig modell for brukere som ønsker veldig ordrette transkripsjoner, men den er nok ikke like nøyaktig som usm og nb-whisper-large-verbatim-bokmål.

nb-wav2vec2-nynorsk

Som for bokmål ser vi en del enkeltbokstaver i nøkkelordsanalysen. Vi ser også tegn til at modellen har gjort andre valg enn i gullstandarden der det er normvariasjon. De automatiske transkripsjonene ser ut til å inneholde færre a-infinitiver enn gullstandarden, blant annet. Dette er ikke egentlige feil, så det er sannsynlig at WER hadde vært noe lavere dersom vi hadde kontrollert for normvariasjon.

Nøkkelordsanalysen viser videre at modellen produserer noen bokmålsord, som *bare*, *mulig*, og *fortsatt*. Stikkprøvene viser at modellen ganske frekvent produserer uttalenære feilstavinger, som *bibern* for *bibelen* og *skjylvfølelsen* for *skuldfølelsen*. Det er også en del bokmålsord i stikkprøvene. I en stor andel av setningene vi har sett på, er transkripsjonen såpass langt fra den normerte skrivemåten at det er vanskelig å tyde hva som ble sagt med utgangspunkt i den. Denne modellen har nok begrensa verdi som transkripsjonsverktøy til nynorsk.

openai-whisper-large-bokmål

Nøkkelordsanalysen viser at diskurspartikler som *liksom*, *altså*, *da*, *så* er mer frekvente i gullstandarden enn i transkripsjonene fra modellen. Dette kommer av at modellen ikke transkriberer ordrett, for det er typisk slike ord som droppes i ikke-ordrette transkripsjoner. Nøkkelordsanalysen viser også at transkripsjonene inneholder nynorskordene *ikkje, ein* og *eg*, selv om vi har spesifisert at språket skal være bokmål.

Transkripsjonene fra denne modellen har store og små bokstaver og tegnsetting og er ikke alltid ordrette. De er gjerne noe kortere enn gullstandarden og dropper ofte gjentakelser,

diskurspartikler og andre ord som kjennetegner muntlig tale. Dette blir ofte mer lesbar tekst, som i (11).

(11)

lydfil:
Verdiborsen_Nummer_59_del_2_030119_228000_618000_s320620_e328280.wav
gullstandard: ***Jaha, nei den*** *det er vanskelig,* ***men*** *det* ***som*** *jeg tenker på når jeg ser* ***de her*** *navnene,* ***men da er det jo da*** *er vi jo på ungdomsskolen.*
openai-whisper-large-bokmål: *Det er vanskelig. Det jeg tenker på når jeg ser navnene,* ***er at*** *vi er på ungdomsskolen.*
WER: 48,28%
SemDist: 0,05

Vi ser imidlertid også at slike ikke-ordrette transkripsjoner en del ganger ikke gjengir meningen på en god måte, og av og til ikke er forståelige i det hele tatt. Det er vanskelig å si om dette er forårsaka av dialektuttale, men i noen av stikkprøvene mistenker vi det.

Feilanalysen viser at modellen av og til hallusinerer, det vil si produserer sekvenser av ord som overhodet ikke stemmer overens med det som sies. For eksempel produserer modellen i et tilfelle *Nå er det slik at vi kan ta det på en annen måte* for et segment der noen kun sier *ja.* I eksempel (12)[15] skriver modellen *unge* 148 ganger.[16]

(12)

lydfil:
Filmpolitiets_podkast_To_ferske_krimtips_fra_Los_Angeles_120522_2130000_2683000_s311241_e314102.wav
gullstandard: *%foreign, nei The Unbearable Weight of Massive Talent.*
openai-whisper-large-bokmål: *Og den unge, un*
WER: 1875%
SemDist: 0,79

Hallusineringer fører gjerne til forholdsvis høy WER og SemDist. Vi har nevnt tidligere at *openai-whisper-large-bokmål* er godt likt av brukere, men ikke får spesielt gode resultater i

[15] Eksempel (12) inneholder et utenlandsk ord, merka som *%foreign* i gullstandarden, og et engelsk navn, så det er å forvente at talegjenkjenningen ikke er spesielt god for akkurat dette segmentet. Det fins imidlertid nyttigere måter å feile på enn å skrive et ikke relatert ord 148 ganger.
[16] Så vidt vi kan se, produserer modellen slike repeterte ord i to tilfeller, så slik ekstrem hallusinasjon ser ikke ut til å være veldig frekvent.

våre tester. Hallusineringene bidrar trolig til at den gjennomsnittlige scoren for WER og SemDist er forholdsvis høy. Man kan se for seg at brukere er villige til å akseptere noen grad av hallusinering, så lenge modellen ellers transkriberer godt, og at det er del av forklaringa på det manglende samsvaret mellom opplevd og målt kvalitet.

openai-whisper-large-nynorsk

Nøkkelordsanalysen viser at transkripsjonene inneholder mange bokmålsord, ord på andre språk, f.eks. *það*, og hyppige sekvenser av tegn som ikke er norske bokstaver. For eksempel forekommer følgende sekvens 154 ganger:
บษษ
ษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษษ
Dette tyder på at modellen hallusinerer svært mye.

Stikkprøvene og feilanalysene bekrefter også at det er mye hallusinering. Når modellen ikke hallusinerer, skriver den ofte bokmål eller dialektnært. Det ser ut som det er noe mer nynorsk når taleren snakker en nynorsknær dialekt, mens for eksempel folk fra østlandsområde har en tendens til å bli transkribert på bokmål eller med dialektord. Feilanalysene viser store mengder hallusinering, ofte med tegn som ikke hører til i det norske alfabetet. Analysene viser at OpenAI Whisper på det nåværende tidspunktet ikke kan brukes til å transkribere til nynorsk.

nb-whisper-large-bokmål

Denne modellen transkriberer i samme stil som openai-whisper-large-bokmål: ikke-ordrett, med store og små bokstaver og tegnsetting. I stikkprøvene transkriberer modellen godt. Selv om den ikke er helt ordrett, ser transkripsjonene ut til å ligge nært opp til det som faktisk sies. Ord og fraser som er typiske for muntlig språk, som *ja, altså* i eksempel (13), kuttes typisk ut.

(13)

> **lydfil:** Utakt_Cola_nummer_350_og_pensjonister_pa_fly_150622_502000_1000000_s365850_e368600.wav
> **gullstandard:** ***Ja, altså*** *ville det tatt seg ut hvis du gikk med tannkost?*
> **nb-whisper-large-bokmål:** *Ville det tatt seg ut hvis du gikk med tannkost?*
> **WER:** 16,67%
> **SemDist:** 0,03

Av og til ser modellen ut til å kutte meningsfulle ord og fraser, som i (14):

(14)

> **lydfil:** Lordagsradet_Mikkel_Niva_Pia_Tjelta_Thomas_Giertsen_230422_574000_949000_s276916_e279510.wav
> **gullstandard:** *Ja,* ***stabilt sideleie,*** *egentlig ganske likt Pia.*
> **nb-whisper-large-bokmål:** *Ja, egentlig ganske likt Pia.*
> **WER:** 28,58%
> **SemDist:** 0,22

I feilanalysene så vi mange segmenter med overlappende tale som ikke hadde korrekt transkripsjon. Blant segmentene med aller høyest WER var det noen tilfeller av hallusinering, som i (15), men det ser ikke ut til å forekomme like hyppig som fra OpenAIs modell.

(15)

lydfil: Sorgens_kapittel_Vanessa_Rudjord_271021_92000_866000_s148990_e149520.wav
gullstandard: *Å!*
nb-whisper-large-bokmål: *Håper du likte denne episoden. Ha det bra!*
WER: 800%
SemDist: 0,73

Alt i alt ser denne modellen ut til å transkribere godt og konsekvent, og transkripsjonene vil fungere godt til for eksempel undertekst og møtetranskripsjon.

nb-whisper-large-nynorsk

Nøkkelordsanalysen viser at modellen i stor grad produserer e-infinitiver, mens gullstandarden har a-infinitiver. Om vi kontrollerte for normvariasjon, ville derfor trolig ordfeilraten vært en del lavere enn 29,54%. Vi ser imidlertid også at modellen produserer en del bokmålsord, som *begynne*, *ikke* og *hva*.

Stikkprøvene viser at modellen ofte produserer korrekt nynorsk, også i tilfeller der taleren ikke snakker nynorsknært. Vi ser imidlertid en del tilfeller av bokmålsord og -bøying i transkripsjonene, som i (16), der modellen ikke har korrekt bøying for *ungdommar* og *kontaktar.* Modellen produserer også *beveger* der gullstandarden har *bevegar*, men siden begge former er tillatt på nynorsk, er ikke dette en egentlig feil. Det samme gjelder for *kvar* og *kor*. Merk også at modellen korrekt skriver *ein*, mens lydopptaket viser at taleren faktisk sier *man*, som ikke ville vært riktig på nynorsk.

(16)

lydfil: Distriktsnyheter_Sorlandet_140922_0_350000_s27030_e37150.wav
gullstandard: *Ungdommar og ekspertar er skeptiske til ny Snapchat-funksjon der ein kan sjå kvar kontaktane dine bevegar seg.*
nb-whisper-large-nynorsk: ***Ungdommer*** *og ekspertar er skeptiske til ny Snapchat-funksjon, der ein kan sjå* ***kor kontakta*** *dine* ***beveger*** *seg.*
WER: 23,53%
SemDist: 0,02

Noen ganger er også hele transkripsjonen på bokmål. Det virker som det er større sjanse for bokmål i transkripsjonene når talerens dialekt er bokmålsnær.

På tross av en del bokmål i transkripsjonene, ser denne modellen ut til å produsere mye god nynorsk, og vi tror den kan være nyttig som et transkripsjonsverktøy for nynorsk.

nb-whisper-large-semantic-bokmål

NB Whispers semantic-modell forkorter og omskriver mer enn nb-whisper-large-bokmål. Ellers ser denne modellen ut til å håndtere dialektvariasjon på en tilsvarende god måte. Forkortingene og omskrivingene er i mange tilfeller ikke betydningsendrende, for eksempel i (17):

(17)

lydfil:
To_i_campingstol_Trine_Skei_Grande_og_Guri_Melby_del_2_070720_93000_530000_s113210_e123660.wav
gullstandard: ***Ja, jeg tror du må huske på at sånn som,*** *jeg antar* ***jo*** *at både Sveinung og Abid har* ***jo*** *vært i en* ***sånn*** *prosess lenge, der man har tenkt* ***litt*** *på hva man ønsker.*
nb-whisper-large-semantic-bokmål: *Jeg antar at både Sveinung og Abid har vært i en prosess lenge der man har tenkt på hva man ønsker.*
WER: 40%
SemDist: 0,05

Imidlertid er det relativt hyppig at forkortingene fjerner såpass mye at det er betydningsendrende, som i (18).

(18)

lydfil:
Gratulerer_med_dagen_17_mai_2019_kl_1000_170519_458000_884000_s226610_e233170.wav
gullstandard: ***Og*** *dronninga er* ***jo*** *høye beskytter for Den norske opera og ballett****, så jeg har vært så heldig å få treffe henne noen ganger i den sammenheng****.*
nb-whisper-large-semantic-bokmål: *Dronningen er høye beskytter for Norsk Operaballett.*
WER: 85,19%
SemDist: 0,25

I feilanalysen er det mange segmenter med overlappende tale. Vi ser også noen tilfeller av hallusinering, men det er uklart om det er mer eller mindre hyppig enn fra nb-whisper-large-bokmål.

Dette ser ut til å være en robust modell som takler variasjon godt. Det kan kanskje være nyttig for en del brukere med de noe kortere transkripsjonene denne modellen produserer, for eksempel for å få plass til undertekst på skjermen. Siden modellen har en tendens til også å kutte meningsfull informasjon, kan det imidlertid hende at det er et tryggere valg for mange å bruke nb-whisper-large-bokmål.

nb-whisper-large-semantic-nynorsk

Nynorsktranskripsjonene har mange av de samme egenskapene som nb-whisper-large-nynorsk: e-infinitiv istedenfor a-infinitiv, noe bokmålsord istedenfor nynorskord, men i stor grad korrekt nynorsk.

nb-whisper-large-verbatim-bokmål

Denne modellen produserer ordrette transkripsjoner med små bokstaver, uten tegnsetting. Transkripsjonene likner en del på de man får fra nb-wav2vec2-bokmål og usm, men vi ser i mindre grad feilstava ord enn fra disse modellene. Transkripsjonene er ofte helt korrekte, og modellen ser ut til å håndtere dialekt godt. Det forekommer imidlertid nynorskord i transkripsjonene. Repeterte ord transkriberes, og avbrutte ord transkriberes også, så når noen avbryter seg selv, får man av og til noen korte bokstavsekvenser som ikke utgjør et ord, men som svarer godt til uttalen. Vi ser ingen tegn til hallusinering, men som alle modellene sliter også denne med overlappende tale.

Som vi så over, har denne modellen den laveste gjennomsnittlige WER-verdien, som tyder på ordrett transkripsjon av høy kvalitet. Stikkprøvene og feilanalysene bekrefter dette inntrykket: Denne modellen ser ut til å ha de mest nøyaktige ordrette transkripsjonene av modellene vi har testa.

nb-whisper-large-verbatim-nynorsk

Verbatim-modellen er også i stand til å produsere gode, ordrette nynorsktranskripsjoner. Som for de andre NB Whisper-modellene produserer modellen i hovedsak e-infinitiver, i motsetning til gullstandarden, så WER-verdien ville nok vært lavere om vi kontrollerte for normvariasjon. Vi ser noen bokmålsord og -former i transkripsjonene. Det virker som bokmål særlig forekommer ved talere av bokmålsnære dialekter.

Utvikling over tid

3 a) På hvilken måte har ny modellarkitektur påvirket kvaliteten på norsk talegjenkjenning?

I kapittel 3 beskreiv vi tre arkitekturer for talegjenkjenningsmodeller: det vi kalte klassiske talegjenkjenningssystemer, wav2vec2 og Whisper. Klassiske talegjenkjenningssystemer er modulære systemer som bruker en akustisk modell, et uttaleleksikon og en språkmodell. Den akustiske modellen trenes på ordrette data, og transkripsjonene som produseres, er også ordrette. Wav2vec2-modeller er transformer-baserte modeller som delvis er trent på taledata uten transkripsjon og delvis på ordrett transkriberte taledata. De transkriberer ordrett, og kan produsere skrivefeil. Whisper-modeller er også transformer-baserte modeller. De kan trenes på ikke-ordrett transkriberte taledata, og transkripsjonene fra modellene trenger heller ikke være ordrette.

Vi har wav2vec2-modeller og Whisper-modeller i vårt utvalg. Vi vet ikke hva slags type arkitektur som er brukt i azure og gcloud-long. Det er imidlertid ikke usannsynlig at de bruker en eller annen variant av det vi har kalt klassiske talegjenkjenningssystemer. Stikkprøvene viser nemlig at de typisk produserer normerte ord når de feiler, ikke feilstava ord. Det kan tyde på at et uttaleleksikon inngår i transkripsjonsprosessen.

Nb-wav2vec2-bokmål og usm, som har en arkitektur som likner en del på wav2vec2, har bedre ordfeilrate og tegnfeilrate enn azure og gcloud-long, men blir slått av azure når vi måler SemDist. Usm framstår som mer nøyaktig i enn azure og gcloud-long i stikkprøvene vi gjorde. Basert på dette utvalget av modeller, er det vanskelig vise at wav2vec2 gir betydelig mye bedre talegjenkjenning på norsk enn klassiske talegjenkjenningssystemer. En fordel med wav2vec2 er imidlertid at kildekoden er åpen, og det er relativt lett å trene nye modeller på åpne datasett. Vi tror at dette er en nyttig arkitektur for akademia, fordi det går an å lage fungerende prototyper på talegjenkjenningmodeller med begrensede ressurser (se Solberg et al. (2023b) for et eksempel på dette).

Whisper-teknologien har hatt betydelig innvirkning på norsk talegjenkjenning. Selv om testene våre ikke gir spesielt gode resultater for openai-whisper-large-bokmål, har den blitt tatt i bruk med stort hell av mange etter den ble lansert i 2022, og brukerne er godt fornøyd med den. Vi tror at en viktig fordel for mange brukere med Whisper-modeller sammenlikna med andre modeller, er at de ikke transkriberer helt ordrett. Det gir transkripsjoner som er lettere å lese og som ser mindre maskingenererte ut. NB Whisper-modellene er foreløpig ikke tatt i bruk av så mange, men de har påviselig høy kvalitet og håndterer dialektvariasjonen på norsk godt.

3 b) På hvilken måte har data fra språkbanken påvirket kvaliteten på norsk talegjenkjenning?

Det er vanskelig å isolere bidraget Språkbankens datasett har hatt på norsk talegjenkjenning. Vi vet ikke hvilke data de kommersielle modellene er trent på, og vi har heller ikke modellpar der den eneste forskjellen er at den ene modellen er trent på data fra Språkbanken og den andre ikke. Noe kan vi likevel si om bidraget til Språkbanken og Nasjonalbiblioteket basert på disse testene.

For det første kan vi konkludere med at det er mulig å trene fungerende talegjenkjenningssystemer for spontan tale kun på åpne datasett fra Språkbanken. Nb-wav2vec2 er nemlig kun trent på åpne datasett i Språkbankens ressurskatalog.

For det andre ser vi en betydelig forbedring i kvalitet fra OpenAI Whisper til NB Whisper. Disse modellene bruker samme arkitektur. Den viktigste forskjellen er hvilke data de er trent på. OpenAI Whisper er trent på lyd og undertekst fra videoer høsta fra internett, inkludert videoer på norsk. NB Whisper er trent på det utvida stortingskorpuset, et datasett i Språkbanken, i tillegg til data fra NRK og lydbøker som Nasjonalbibliotekets AI-lab har fått tilgang til. Det er dermed i all hovedsak data fra Språkbanken og data samla inn av Nasjonalbiblioteket som er årsak til kvalitetsforbedringen. Det er imidlertid viktig å understreke at det kun er datasettet fra Språkbanken som er allment tilgjengelig og som dermed kan brukes av andre som ønsker å utvikle talegjenkjenning. NB Whisper er imidlertid tilgjengelig med åpen kildekode og kan videretrenes ytterligere av andre uten restriksjoner, så brukere og utviklere kan få glede av dataene selv om ikke alle datasettene er åpent tilgjengelige.

6. Konklusjon

I innledninga formulerte vi disse tre hovedmålene for rapporten:

1. Evaluere status for norsk talegjenkjenning
2. Finne ut hva slags feil norske talegjenkjenningssystemer gjør og hva Språkbanken skal bruke tid og ressurser på videre
3. Vise forbedringene i norsk talegjenkjenning over tid

Som konklusjon vil vi oppsummere de funnene vi har gjort og knytte dem opp mot disse målene.

Vår testing viser at det fins velfungerende talegjenkjenningssystemer for bokmål og nynorsk (mål 1). Både kvantitative og kvalitative analyser vi har gjennomført, viser at systemene fra NB Whisper nok er de beste for bokmål, og utvilsomt best for nynorsk. Vi har vist at både transformer-teknologien og økt tilgang på relevante data har bidratt betydelig til at norsk talegjenkjenning i dag fungerer forholdsvis godt (mål 3).

Det er imidlertid noen punkter der vi ser potensial for forbedring (mål 2):

1. **Bokmål og nynorsk**: I transkripsjonene som støtter begge målformer, ser vi en del tilfeller av bokmålsord og -former i nynorskranskripsjoner og nynorskord og -former i bokmålstranskripsjoner. Dette problemet ser ofte ut til å bli framkalt av bokmålsnære og nynorsknære dialekter. Vi mener det kan være en god idé å undersøke hvordan målformene kan skilles bedre fra hverandre, enten ved å rydde i treningsdataene, skape nye treningsdata eller videreutvikle selve talegjenkjenningssystemene.
2. **Overlappende tale:** Alle systemene vi har testa, sliter med å transkribere korrekt når det er overlappende tale. Dette er ikke overraskende. Overlappende tale er også utfordrende for mennesker. Imidlertid er det viktig at systemer som skal håndtere dialoger og spontan tale, håndterer overlappende tale. Dette er et område det bør forskes mer på. Talegjenkjenning av dialoger er et av temaområdet for SCRIBE-prosjektet, som Nasjonalbiblioteket deltar i.
3. **Dialekter:** Vi har observert at systemene er bedre på noen dialekter enn andre. Dette gjelder særlig for nynorsktranskripsjon, der nynorsknære dialekter blir transkribert bedre enn bokmålsnære. Vi ser tilsvarende tendenser på bokmål, selv om de beste modellene ikke ser ut til å bli like sterkt påvirka av dialekt i vårt utvalg. Forskning på talegjenkjenning av norske dialekter er et annet temaområde for SCRIBE-prosjektet, og Nasjonalbiblioteket og Språkrådet bør følge nøye med på resultatene derfra.
4. **Ordrett og ikke-ordrett transkripsjon:** Det fins systemer av høy kvalitet som transkriberer ordrett og ikke-ordrett på norsk. Vi tror at det er bruksområder for begge. Ordrette transkripsjoner egner seg best i sammenhenger der nøyaktighet er viktig, for eksempel i transkripsjon av avhør og i lingvistisk forskning. Ikke-ordrett transkripsjon egner seg trolig bedre for møtereferater, undertekst o.l., der det er viktig med god skriftlig norsk. Vi ser imidlertid at systemer som produserer ikke-ordrett transkripsjon, kan fjerne for mye av det som opprinnelig ble sagt, slik at viktige aspekter ved meninga forsvinner. Vi tror det trengs mer forskning på ikke-ordrett

transkripsjon, både for å finne ut av hvordan man kan unngå at viktig informasjon forsvinner, hvordan man skal teste ikke-ordrette systemer på en god måte, hva brukere opplever som gode transkripsjoner og når de opplever at transkripsjonen er for fjern fra det som faktisk sies.

5. **Hallusinering og fremmede språk:** Systemer som er trent på flere språk samtidig, spesielt USM og OpenAI Whisper, har av og til en tendens til å transkribere på galt språk. Ikke-ordrette systemer har i tillegg en tendens til å hallusinere. Vi har imidlertid ikke observert spesielt mye hallusinasjon fra NB Whisper, selv om det av og til forekommer.
6. **Kjønn:** De fleste systemene gjør det litt bedre på kvinnestemmer enn på mannsstemmer. Det er uklart hvorfor det er sånn, men det er rapportert om tilsvarende effekter for andre språk (jf. Feng et al, 2024 og referanser der). Vi tror ikke denne forskjellen er stor nok til at det betyr så mye i praksis, men det kan være et interessant område for videre forskning.

I denne rapporten har vi testa på et testsett med kringkastingsopptak. Dette testsettet har stor grad av variasjon av sjanger, alder, kjønn, dialekt og opptaksforhold. Det er imidlertid en del former for variasjon som ikke fanges opp ved å teste på dette materialet. Testsettet inneholder for eksempel i liten grad barn, andrespråkstalere og personer med språkvansker. Disse gruppene er trolig også underrepresentert i treningsdataene til de fleste modellene, og det er sannsynlig at systemene ikke håndterer dem like godt som voksne mennesker uten språkvansker som er født i Norge. Samtidig er det viktig at teknologien fungerer godt også for disse gruppene. Vi tror det ville være en god idé å gjøre noen mindre tester på tale fra slike underrepresenterte grupper.

Som vi har påpekt flere ganger, samsvarer ikke de tilgjengelige kvalitetsmålene alltid med opplevd kvalitet. Ikke-ordrette systemer er særlig vanskelige å teste automatisk. For eksempel er mange brukere fornøyd med OpenAIs Whisper på bokmål, og i vår kvalitative testing ser vi at dette systemet ofte produserer gode transkripsjoner, men det får forholdsvis høye WER- og SemDist-verdier. For å få mer innsikt i hvor godt de tilgjengelige kvalitetsmålene samsvarer med hvor fornøyd brukere er, bør man gjennomføre brukertesting. Det kan enten gjøres ved hjelp av eksperimenter eller ved å legge til en tilbakemeldingsfunksjon i transkripsjonsverktøy som har mange brukere.

Referanser

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems,* 33, s. 12449-12460.

Feng, S., Halpern, B. M., Kudina, O., & Scharenborg, O. (2024). Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84, 101567.

Kim, S., Arora, A., Le, D., Yeh, C. F., Fuegen, C., Kalinli, O., & Seltzer, M. L. (2021). Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. *arXiv preprint* arXiv:2104.02138.

Ljubešić, N., Koržinek, D., Rupnik, P., & Jazbec, I. P. (2022). ParlaSpeech-HR-a freely available ASR dataset for croatian bootstrapped from the parlaMint corpus. I *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation Conference*, s 111-116.

Pojanapunya, P., & Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), s. 133-167.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. I Proceedings of the International Conference on Machine Learning, s. 28492-28518.

de la Rosa, J., Braaten, R. A., Kummervold, P. E., & Wetjen, F. (2023). Boosting Norwegian Automatic Speech Recognition. I *Proceedings of NoDaLiDa 2023*, s. 555-564

Solberg, P. E., Beauguitte, P., Kummervold, P. E., & Wetjen, F. (2023). A Large Norwegian Dataset for Weak Supervision ASR. I *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023),* s. 48-52.

Solberg, P. E., & Ortiz, P. (2022). The Norwegian Parliamentary Speech Corpus. I *Proceedings of LREC 2022*, s 1003-1008.

Solberg, P. E., Ortiz, P., Parsons, P., Svendsen, T., & Salvi, G. (2023). Improving generalization of Norwegian ASR with limited linguistic resources. I *Proceedings of LREC 2023*, s. 508-517.

Stolcke, A., & Droppo, J. (2017). Comparing human and machine errors in conversational speech transcription. *arXiv preprint* arXiv:1708.08615.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... & Wu, Y. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint* arXiv:2303.01037.