Transcriptomatic

Författare: Anton Lundström, Robert Zeijlon

Datum: 16 maj 2025

Ämne: Utvärdering av Speech-to-Text-modeller för svenska

Sammanfattning (Abstract)

Till skillnad från engelska är utbudet av högkvalitativa **Speech-to-Text (STT)** modeller för svenska mer begränsat. Detta gör det särskilt viktigt att noggrant utvärdera befintliga modellers prestanda för att identifiera de som bäst hanterar det svenska språket. Vårt test fokuserar specifikt på hur väl olika STT-modeller hanterar längre ljudformat. Detta är avgörande för att bedöma deras lämplighet i scenarier som möten, intervjuer och andra situationer där talet ofta är mer komplext och sammanhängande. Genom att utvärdera modellerna på denna typ av data siktar vi på att sammanställa en bild av hur pass bra de är på att transkribera verkliga samtal på svenska.

Förkortningar

STT - Speech-to-Text, **WER** - Word Error Rate, **CER** - Character Error Rate, **BERTScore** - ett mått på textlikhet, **METEOR** - ett mått på maskinöversättningskvalitet, **Podman** - ett containerhanteringsverktyg

---

Innehållsförteckning

---

1. Introduktion

1.1 Bakgrund

Det finns många modeller som är väldigt bra på att transkribera engelska, både betal- och open source-modeller. Problemet är att det på svenska inte finns lika många högkvalitativa alternativ. Därför vill vi testa och utvärdera hur bra befintliga modeller presterar på svenska. För att göra detta behövde vi skapa lite egen data. Vi har kontaktat några poddare för att få tillåtelse att använda deras material som data och har även hämtat material från Riksdagen som är fritt att använda för sådana här syften. Denna data har vi sedan manuellt transkriberat. Vi är särskilt intresserade av hur väl **KBLabs** modeller presterar.

Sammanfattning: KBLab KB-Whisper KBLab har utvecklat KB-Whisper, en ny tal-till-text-modell speciellt framtagen för svenska. Den bygger på OpenAIs populära Whisper-modell men har finjusterats på en massiv datamängd om 50 000 timmar svenskt tal (från bland annat TV-textning och riksdagsprotokoll). Resultatet är en kraftig förbättring av prestandan för svensk taligenkänning. Tester visar en genomsnittlig minskning av ordfelet (**WER**) med hela 47 % jämfört med OpenAIs bästa modell (Whisper-large-v3). En stor fördel är att även mindre KB-Whisper-modeller presterar bättre än OpenAIs större modeller. Detta gör högkvalitativ svensk taligenkänning mer tillgänglig och kostnadseffektiv. Modellerna är fritt tillgängliga via KBLab på HuggingFace.

1.2 Problemformulering

Trots ett ökat behov av automatiserad tal-till-text-konvertering på svenska, är utbudet av högpresterande **STT**-modeller för språket begränsat jämfört med engelska. Detta skapar en kunskapslucka kring vilka modeller som bäst lämpar sig för olika applikationer, särskilt när det gäller transkription av längre, mer komplexa svenska ljudinspelningar som poddar och riksdagsdebatter. Ett otillräckligt underlag för att välja optimal modell kan leda till ineffektiva processer, ökade kostnader för manuell transkribering, och sämre kvalitet på textdata. Det är därför viktigt att systematiskt utvärdera befintliga **STT**-modellers förmåga att hantera svenska taldataspecifika utmaningar.

1.3 Syfte och Frågeställningar

Syftet med denna rapport är att utvärdera och jämföra prestandan hos ett urval av open-source och kommersiella **Speech-to-Text**-modeller för svenska, med särskilt fokus på deras förmåga att transkribera långformat ljud. Vi kommer att besvara följande forskningsfrågor:

1. Hur skiljer sig olika **STT**-modellers transkriptionskvalitet, mätt med standardmetriker som **WER**, **CER**, **BERTScore** och **METEOR**, för svenskt långformat ljud?
2. Vilka modeller uppvisar bäst prestanda när det gäller kontextuell korrekthet och övergripande förståelse av svenskt tal i längre inspelningar, baserat på AI-baserad utvärdering?
3. Vilka implikationer har resultaten för valet av **STT**-modeller för svenska i praktiska tillämpningar, och vilka potentiella framtida utvecklingsområden kan identifieras?

1.4 Avgränsningar

Utvärderingen kommer att inkludera specifika open-source- och kommersiella modeller och fokusera på långformat ljud på svenska från poddar och Riksdagens inspelningar. Vi kommer inte att inkludera specialiserade dialekter. Datan är även begränsad i storlek, vilket kan påverka generaliserbarheten av resultaten.

1.5 Disposition

Denna rapport är strukturerad enligt följande: Avsnitt 1 introducerar bakgrunden, problemformuleringen, syftet och avgränsningarna för studien. Avsnitt 2 presenterar den teoretiska referensramen och en genomgång av relevant litteratur. Avsnitt 3 beskriver den metod som använts för datainsamling, urval och analys. Avsnitt 4 presenterar resultaten av modellutvärderingen. Avsnitt 5 innehåller en diskussion av resultaten, deras tolkning, jämförelse med tidigare forskning, samt en metoddiskussion och identifierade implikationer. Slutligen, i Avsnitt 6, dras slutsatser och förslag på framtida arbete ges.

---

2. Teoretisk Referensram och Litteraturgenomgång

2.1 Relevant Teori/Modeller

Tal-till-text-teknologi (Speech-to-Text, STT) bygger på komplexa algoritmer inom artificiell intelligens och maskininlärning. Grundläggande för **STT**-system är akustisk modellering, språkmodellering och avkodning.

- **Akustisk modellering:** Denna komponent omvandlar ljudvågor till fonetiska representationer eller sekvenser av sub-ord-enheter. Historiskt har Dolda Markov-modeller (HMM) varit dominande, men på senare tid har djupa neurala nätverk (**DNN**), särskilt Recurrent Neural Networks (**RNN**), Convolutional Neural Networks (**CNN**) och Transformer-arkitekturer, revolutionerat området. Dessa modeller kan lära sig komplexa mönster i taldata, inklusive variationer i uttal, ton och hastighet.
- **Språkmodellering:** Efter den akustiska analysen används en språkmodell för att förutsäga sannolikheten för ordsekvenser. Detta hjälper till att disambiguera fonetiskt lika ljud och säkerställa att transkriptionen är grammatiskt korrekt och meningsfull. Moderna språkmodeller är ofta baserade på stora textkorpusar och utnyttjar neurala nätverk som LSTMs (Long Short-Term Memory) eller Transformer-baserade modeller (t.ex. GPT-familjen) för att förstå kontext och semantik.
- **Avkodning:** Avkodaren kombinerar information från den akustiska modellen och språkmodellen för att hitta den mest sannolika ordsekvensen som matchar ljudingången.

En central utveckling de senaste åren är användningen av **end-to-end-modeller**, som exempelvis OpenAIs **Whisper**. Dessa modeller hanterar hela processen från rått ljud till text i ett enda neuralt nätverk, vilket förenklar arkitekturen och ofta förbättrar prestandan genom att optimera alla komponenter gemensamt. Whisper-modellen utmärker sig genom sin utbildning på en mycket stor och diversifierad datamängd bestående av "weakly supervised" ljuddata (ljud-text-par där texten inte är perfekt alignerad) från webben, vilket gör den robust för olika accenter, bakgrundsljud och ämnen.

KBLab KB-Whisper är ett exempel på en finjusterad version av Whisper för ett specifikt språk – svenska. Genom att träna Whisper-modellen på en omfattande korpus av svenskt tal (50 000 timmar) har KBLab kunnat anpassa modellen för att bättre hantera svenska fonem, ordföljd och grammatiska strukturer, vilket leder till en signifikant minskning av ordfelet jämfört med den generiska Whisper-modellen.

2.2 Tidigare Forskning

Tidigare forskning inom **STT** för svenska har ofta fokuserat på att anpassa generella modeller eller bygga modeller från grunden med mindre, språkspecifika datamängder. Historiskt sett har prestandan för svenska **STT**-system legat efter den för engelska på grund av mindre tillgång till stora, högkvalitativa träningsdata. Studier har visat att framsteg inom maskininlärning, särskilt djupa neurala nätverk, har bidragit till betydande förbättringar även för mindre språk.

Forskning kring utvärdering av **STT**-modeller betonar ofta vikten av att använda relevanta metriker utöver de traditionella **WER** och **CER**. Medan **WER (Word Error Rate)** och **CER (Character Error Rate)** ger en kvantitativ bild av fel på ord- respektive teckennivå, fångar de inte alltid den semantiska korrektheten eller den kontextuella förståelsen. Därför har mått som **BERTScore** och **METEOR** blivit alltmer relevanta. **BERTScore** utnyttjar förtränade språkmodeller (som BERT) för att bedöma den semantiska likheten mellan en predikerad och en referenstext, vilket ger en mer nyanserad bild av transkriptionens kvalitet. **METEOR**, å andra sidan, tar hänsyn till synonymer och parafraser, vilket ger en mer "mänsklig" bedömning av översättningskvalitet, även om den ursprungligen utvecklades för maskinöversättning, är den applicerbar på STT-utvärdering.

Den begränsade tillgången på stora, transkriberade svenska taldatamängder har länge varit en utmaning för att utveckla robusta **STT**-modeller. Projekt som **KBLabs** arbete med **KB-Whisper** visar på en viktig trend: att utnyttja stora mängder "weakly supervised" data (t.ex. TV-textning och Riksdagsprotokoll) i kombination med finjustering är en effektiv strategi för att förbättra **STT**-prestanda för språk med mindre resurser. Forskning indikerar att modellernas prestanda varierar avsevärt beroende på ljudkvalitet, bakgrundsljud, talarens dialekt och tals hastighet, vilket understryker behovet av omfattande utvärderingar på realistiska datamängder.

---

3. Metod

Utvärderingen av **Speech-to-Text**-modellerna genomfördes genom en kombination av tekniska implementeringar och systematiska mätningar, med fokus på att jämföra prestandan hos ett urval av open-source och kommersiella modeller för svenska långformat ljud. Processen involverade datainsamling, utveckling av ett utvärderingsskript samt beräkning av både standard **STT**-metriker och AI-baserade kontextuella likhetsmått.

3.1 Val av Metod

Denna studie använde en kvantitativ utvärderingsmetod där **STT**-modellernas transkriptionskvalitet mättes objektivt med hjälp av olika metriker. Detta kompletterades med en kvalitativ, AI-baserad bedömning av kontextuell korrekthet och övergripande förståelse för att få en mer nyanserad bild av modellernas prestanda. Valet att inkludera både open-source och kommersiella modeller möjliggör en bred jämförelse av tillgängliga lösningar.

3.2 Datainsamling

Vi samlade in ett eget, manuellt transkriberat dataset. Detta dataset bestod av cirka 5 timmar svenskt långformat ljud, hämtat från poddavsnitt (efter inhämtat tillstånd från poddare) och debatter från Riksdagen (som är fritt tillgängligt). Ljudfilerna transkriberades manuellt för att skapa "ground truth"-text som modellernas transkriptioner sedan jämfördes mot. Denna process var tidskrävande men nödvändig för att skapa en utvärderingsbas för svenska, då det finns begränsat med befintliga, högkvalitativa, manuellt transkriberade dataset för den typ av data med svenskt tal.

3.3 Urval

Urvalet av **STT**-modeller för utvärderingen inkluderade:

- **Open-source-modeller:** Huvudsakligen olika varianter av **OpenAI Whisper** (tiny, small, medium, large-v3, large-v3-turbo) och **KBLab KB-Whisper** (base, medium, large, small, tiny). Dessa modeller representerar den senaste utvecklingen inom community-drivna **STT**-lösningar, med KBLabs modeller specifikt tränade för svenska.
- **Kommersiella API:er:** Inkluderar **Azure**, **Deepgram** och **Elevenlabs**. Dessa valdes ut som representanter för ledande kommersiella aktörer som erbjuder **STT**-tjänster med stöd för svenska.

3.4 Analysmetod

För att utvärdera transkriptionskvaliteten användes ett utvärderingsskript som beräknade följande metriker:

- **Word Error Rate (WER):** Mäter antalet felaktiga ord (insättningar, borttagningar, substitutioner) i den transkriberade texten jämfört med referenstexten, dividerat med det totala antalet ord i referenstexten. En lägre **WER** indikerar bättre prestanda.
  - $WER = \frac{(Substitutions + Insertions + Deletions)}{Totala Antalet Ord i Referenstexten}$
- **Character Error Rate (CER):** Mäter antalet felaktiga tecken (insättningar, borttagningar, substitutioner) i den transkriberade texten jämfört med referenstexten, dividerat med det totala antalet tecken i referenstexten. Användbart för att identifiera små stavfel eller fel i språk utan tydliga ordgränser. En lägre **CER** indikerar bättre prestanda.
- **BERTScore:** En semantisk likhetsmetrik som använder embeddings från BERT (Bidirectional Encoder Representations from Transformers) för att mäta hur lik predikerad text är referenstexten, med fokus på betydelse snarare än exakta ordmatchningar. Värden ligger mellan 0 och 1, där närmare 1 indikerar högre semantisk likhet.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):**Även om ursprungligen utvecklad för maskinöversättning, är **METEOR** användbar för **STT**. Den tar hänsyn till exakta ordmatchningar, synonymer och parafraser, vilket ger en mer "mänsklig" bedömning av översättningskvalitet, även om den ursprungligen utvecklades för maskinöversättning, är den applicerbar på STT-utvärdering.

Utöver dessa standardmetriker utförde vi även en **AI-baserad kontextuell likhetsbedömning** med hjälp av modellen `o4-mini` från OpenAI. Denna modell användes för att generera en subjektiv bedömning av kontextuell korrekthet (`llm_eval_kontextuell_korrekthet`) och övergripande förståelse (`llm_eval_overgripande_forstaelse`) för varje transkription. Detta gav en "mänskligare" kvalitetsbedömning som fångar upp fall där små fel inte påverkar textens generella budskap.

För utvärdering av lokala modeller modifierade vi en klient till WhisperLive, ett open-source-program för transkribering av ljuddata som en ström. Vi anpassade den för att ta en ljudfil som input och sedan transkribera den till text, med möjlighet att spara transkriptionen till en fil. WhisperLive-serverkomponenten lämnades omodifierad och hostades med hjälp av **Podman**, ett containerverktyg.

För utvärdering av kommersiella modeller användes deras respektive API:er för att skicka ljudfiler och ta emot transkriptioner. Vi såg medvetet till att stänga av funktioner så som: tids markörer och identifiering av talare. Detta av två anledningar:

1. Det är inte kompatibelt med våra manuella transkriberingar.
2. Dessa verktyg förbättrar generellt prestandan på modellen utan att vara en del av modellen och denna studie utvärderade modeller och inte andra användbara verktyg.

3.5 Etiska Överväganden

[Inkludera etiska överväganden här om det finns]

---

4. Resultat

Resultaten av utvärderingen sammanställs nedan, baserat på de metriker som beskrevs i metodavsnittet. Tabellen visar prestandan för de utvärderade STT-modellerna.

Modell	WER_score	CER_score	BERTScore	METEOR_score	Cleaned_WER_score	Cleaned_CER_score	Cleaned_BERTScore	Cleaned_METEOR_score	llm_eval_sc
KBLab-kb-whisper-base	0.33166	0.20782	0.65729	0.55409	0.24770	0.18781	0.90422	0.57690	0.75
KBLab-kb-whisper-large	0.24175	0.13118	0.66317	0.63536	0.15792	0.11184	0.94549	0.66722	0.81
KBLab-kb-whisper-medium	0.54385	0.44668	0.70394	0.35890	0.47572	0.43043	0.88218	0.37302	0.76
KBLab-kb-whisper-small	0.37814	0.26798	0.65643	0.51627	0.30128	0.25008	0.89158	0.53576	0.75
KBLab-kb-whisper-tiny	0.34656	0.21617	0.65699	0.54031	0.26663	0.19659	0.90162	0.55980	0.67
azure-api	0.28809	0.17719	0.67153	0.61018	0.20710	0.15861	0.92492	0.63928	0.81
deepgram-api	0.39059	0.25575	0.65734	0.50948	0.30033	0.23540	0.86234	0.53217	0.66
elevenlabs-api	0.43961	0.35119	0.65441	0.50893	0.37704	0.33395	0.84943	0.54012	0.67
openAI_large-v3	0.26270	0.12530	0.65663	0.62886	0.15496	0.10029	0.95076	0.67560	0.71
openAI_large-v3-turbo	0.26358	0.12068	0.65997	0.65088	0.15281	0.09498	0.95242	0.69641	0.71
openAI_medium	0.41281	0.30143	0.66374	0.46673	0.34077	0.28446	0.85336	0.48434	0.69
openAI_small	0.37542	0.22882	0.65538	0.52215	0.29475	0.20888	0.89376	0.54638	0.69
openAI_tiny	0.55930	0.29462	0.65677	0.43568	0.48424	0.27082	0.84069	0.44180	0.51

4.1 Resultat relaterat till Frågeställning 1/Tema A

Med avseende på traditionella felkvoter som **WER (Word Error Rate)** och **CER (Character Error Rate)**, framgår det att de större modellerna från både **KBLab** och **OpenAI** generellt presterar bäst. Specifikt har **KBLab-kb-whisper-large** och **openAI_large-v3** samt **openAI_large-v3-turbo** de lägsta **WER**- och **CER**-värdena. Detta indikerar att de är bäst på att transkribera ord och tecken korrekt. Även om **WER**- och **CER**-värdena för de kommersiella API:erna (**Azure**, **Deepgram**, **Elevenlabs**) är högre än de bästa open-source-alternativen, visar **Azure-api** en relativt stark prestanda jämfört med Deepgram och Elevenlabs. Noterbart är att "Cleaned"-versionerna av **WER** och **CER** (som indikerar en efterbearbetad transkription) visar en signifikant förbättring för alla modeller, vilket understryker vikten av efterbearbetning.

När det gäller **BERTScore** och **METEOR**, som mäter semantisk likhet och översättningskvalitet, uppvisar **openAI_large-v3-turbo** de högsta **METEOR**-värdena, vilket tyder på en utmärkt förmåga att fånga betydelsen även när exakta ord inte matchar. **KBLab-kb-whisper-large** presterar också mycket väl i dessa kategorier, vilket bekräftar dess höga kvalitet för svenskt tal. Även här ser vi en förbättring i "Cleaned"-värdena, vilket förstärker bilden av att modellerna upprätthåller den ursprungliga betydelsen även med små fel.

4.2 Resultat relaterat till Frågeställning 2/Tema B

Den **AI-baserade utvärderingen**, representerad av `llm_eval_score`, `llm_eval_kontextuell_korrekthet` och `llm_eval_overgripande_forstaelse`, ger en kvalitativ insikt i modellernas förmåga att förstå och återge kontext korrekt. Här framträder **KBLab-kb-whisper-large** och **Azure-api** som de modeller som får de högsta poängen. Deras höga värden för `llm_eval_kontextuell_korrekthet` (0.8 respektive 0.76) och `llm_eval_overgripande_forstaelse` (0.82 respektive 0.86) indikerar att dessa modeller inte bara transkriberar korrekt på ordnivå, utan också lyckas upprätthålla meningsfullhet och sammanhang i den transkriberade texten. De openAI-modeller som testats har något lägre AI-baserad utvärderingspoäng, trots deras starka prestanda på de mer kvantitativa metriker som **WER** och **CER**. Detta kan bero på att AI-utvärderingen är mer känslig för nyanser i språket och den övergripande läsbarheten. De mindre modellerna, särskilt `openAI_tiny`, har betydligt lägre poäng i denna kategori, vilket förväntas då de är mer begränsade i sin kapacitet.

---

5. Diskussion

5.1 Sammanfattning av Resultat

Vår utvärdering av **Speech-to-Text**-modeller för svenska visar tydligt att det finns en betydande variation i prestanda mellan olika modeller, både open-source och kommersiella. De större modellerna, i synnerhet **KBLab-kb-whisper-large** och **OpenAI_large-v3-turbo**, presterar genomgående bäst när det gäller traditionella felkvoter som **WER** och **CER**. Detta understryker vikten av modellstorlek och träningsdata för hög precision. **KBLabs** framgång med sin finjusterade **KB-Whisper**-modell för svenska är särskilt anmärkningsvärd, då den i många fall matchar eller överträffar OpenAIs generiska modeller för svenskt tal.

När det gäller semantisk korrekthet och kontextuell förståelse, mätt med **BERTScore**, **METEOR** och vår LLM-baserade utvärdering, bekräftas de större modellernas överlägsenhet. De uppvisar högre scores, vilket indikerar att de inte bara transkriberar ord korrekt, utan också lyckas bevara textens ursprungliga mening och sammanhang. **Azure-api** visar en stark prestanda i den AI-baserade utvärderingen, trots något högre **WER/CER** än de allra bästa open-source-alternativen, vilket kan tyda på att dess interna språkmodellering är effektiv på svenska.

De "cleaned" resultaten för **WER** och **CER** visar genomgående en markant förbättring för samtliga modeller. Detta indikerar att även mindre efterbearbetning kan ha en stor inverkan på transkriptionskvaliteten, vilket är relevant för praktiska tillämpningar.

5.2 Tolkning och Analys av Resultat

Resultaten indikerar att det för svenskt långformat ljud är fördelaktigt att välja modeller som är antingen specifikt tränade på svenska (som **KBLabs** modeller) eller mycket stora, generiska modeller (som OpenAIs large-modeller) som har exponerats för en bred uppsättning språk. **KBLabs** framgång med **KB-Whisper** bekräftar hypotesen att språkspecifik finjustering på en stor datamängd ger betydande fördelar för små språk. Att även de mindre **KB-Whisper**-modellerna presterar bra relativt sin storlek är en viktig slutsats för resursbegränsade applikationer.

Den observerade skillnaden mellan "råa" **WER/CER**-värden och "cleaned" värden understryker att **STT**-modeller fortfarande kan ha utmaningar med inkonsekvenser, interpunktion och andra mindre fel som kan korrigeras med efterbearbetning. Detta är särskilt relevant för längre transkriptioner där tystnader kan leda till "hallucineringar" (generering av icke-existerande ord).

Övergången bland kommersiella aktörer, som **Azure** och Google, till LLM-baserad transkribering är en intressant observation. Även om denna teknik har enorm potential för kontextuell förståelse, är den fortfarande ny och kan vara mindre stabil för rena transkriptionsändamål jämfört med dedikerade **STT**-modeller. Våra resultat för **Azure-api** speglar detta: god kontextuell förståelse, men inte alltid de lägsta felkvoterna. Detta kan förklaras av att LLM-baserade system prioriterar att meningen ska vara korrekt även om det inte är en ord-för-ord-matchning.

5.3 Jämförelse med Tidigare Forskning

Våra resultat överensstämmer med tidigare forskning som visar att tillgången till stora och relevanta träningsdata är avgörande för **STT**-prestanda, särskilt för mindre språk. **KBLabs** framsteg med **KB-Whisper** validerar strategin att finjustera generiska modeller som **Whisper** på språkspecifika korpusar för att uppnå toppresultat. Detta bekräftar den generella trenden inom AI att specialisering på språk kan ge avsevärda prestandavinster. Användningen av **BERTScore** och **METEOR** i kombination med **WER/CER** är också i linje med modern utvärderingspraxis som strävar efter att fånga mer än bara exakta ordmatchningar. Vår AI-baserade utvärdering är ett komplement till dessa, vilket ger en mer kvalitativ bedömning av förståelse, liknande hur mänskliga bedömare kan värdera en transkription.

5.4 Metoddiskussion (Styrkor och Svagheter)

Styrkor:

- **Relevans:** Utvärderingen fokuserade på långformat svenskt ljud, vilket är direkt tillämpligt för många praktiska användningsområden som mötesprotokoll, poddtranskriptioner och intervjuer.
- **Bred jämförelse:** Inkludering av både open-source och kommersiella modeller ger en omfattande bild av tillgängliga alternativ.
- **Mångsidiga metriker:** Användningen av **WER**, **CER**, **BERTScore**, **METEOR** samt en AI-baserad utvärdering ger en robust och nyanserad bild av modellernas prestanda, där både exakthet och semantisk korrekthet beaktas.
- **Egen transkriberad data:** Genom att skapa ett eget manuellt transkriberat dataset säkerställdes högkvalitativ "ground truth"-data för svenska, vilket är en bristvara.

Svagheter:

- **Datakvalitet och begränsning:** Den manuellt transkriberade datamängden är relativt begränsad (ca 5 timmar). Detta kan påverka resultatens generaliserbarhet och göra att de inte är helt representativa för all tänkbar svensk taldialekt, ljudkvalitet eller ämnesområde. Manuella transkriptioner är även tidskrävande och kan innehålla mänskliga fel.
- **AI-baserad utvärdering:** Även om den AI-baserade utvärderingen ger värdefulla kvalitativa insikter, är den beroende av den specifika LLM-modellens (o4-mini) förmåga att bedöma transkriptionerna. Dess objektivitet är inte absolut och kan vara subjektivt påverkad av modellens egna "fördomar" eller träningsdata.
- **LLM-baserad transkribering:** Upplevelsen av att kommersiella API:er skiftar till LLM-baserad transkribering under projektets gång indikerar en dynamisk marknad. Detta kan innebära att tekniken är under snabb utveckling och att våra resultat för dessa modeller kan vara föråldrade relativt snabbt.

5.5 Implikationer och Rekommendationer

Resultaten har flera viktiga implikationer för val av **STT**-modeller för svenska:

- **Första val för svenska:** För högkvalitativ transkription av svenska är **KBLabs** finjusterade **KB-Whisper**-modeller ett mycket starkt alternativ, särskilt `kb-whisper-large`. De erbjuder prestanda i nivå med eller över de bästa generiska modellerna, och dessutom som open-source.
- **Större modeller är bättre:** Generellt sett presterar större modeller bättre än mindre, både kvantitativt och kvalitativt. Val av modellstorlek bör balanseras mot tillgängliga beräkningsresurser.
- **Efterbearbetning är kritiskt:** Oavsett modellval kan en efterbearbetningsprocess för att korrigera mindre fel och hallucineringar avsevärt förbättra den slutliga transkriptionskvaliteten. Att implementera mer avancerad meningssegmentering (chunking/splitting) och att hantera tystnader kan förbättra resultaten.

- **Framtida potential för LLM-baserad transkribering:** Även om dagens LLM-baserade transkription fortfarande är under utveckling för rena **STT**-uppgifter, pekar **Azures** resultat på att tekniken har en stark förmåga att förstå kontext. Framtida specialiserade LLM-modeller för transkribering, gärna open-source, kan komma att revolutionera området.
- **Behov av mer data:** För att ytterligare förbättra **STT**-modeller för svenska krävs större, högkvalitativa och diversifierade dataset. Detta kommer troligen att kräva initiativ från forskningsinstitut, myndigheter eller andra intressenter då kommersiella aktörer har begränsade incitament att investera tungt i mindre språk som svenska.

---

6. Slutsatser och Framtida Arbete

6.1 Slutsatser

Denna rapport har utvärderat prestandan hos flera open-source och kommersiella **Speech-to-Text (STT)**-modeller för svenskt långformat ljud. Våra slutsatser pekar på att det finns högkvalitativa alternativ för svensk taligenkänning, med **KBLab-kb-whisper-large** och **OpenAI_large-v3-turbo** som framträdande modeller med låga felkvoter (**WER/CER**) och hög semantisk korrekthet (**BERTScore/METEOR**). **KBLabs** specifika satsning på svenska har resulterat i en modell som är mycket konkurrenskraftig och bekräftar värdet av språkspecifik finjustering. Den AI-baserade utvärderingen belyste att modeller som **KBLab-kb-whisper-large** och **Azure-api** är skickliga på att bevara den kontextuella förståelsen och övergripande budskapet i transkriptionerna. Trots modellernas framsteg, är efterbearbetning av transkriptioner fortfarande en viktig faktor för att uppnå optimal kvalitet, då även de bästa modellerna kan generera mindre fel eller "hallucinera" vid tystnader. Framtiden för **STT**-teknologi tycks alltmer inkludera LLM-baserade metoder, vilket kan leda till ytterligare förbättringar i kontextuell förståelse.

6.2 Framtida Arbete

För att bygga vidare på denna studie och ytterligare förbättra **Speech-to-Text**-lösningar för svenska, föreslås följande områden för framtida arbete:

- **Utökning av Dataset:** Samla in och manuellt transkribera ett betydligt större och mer diversifierat dataset av svenskt långformat ljud. Detta skulle inkludera fler dialekter, varierande ljudkvalitet, olika ämnesområden och inspelningstyper för att öka generaliserbarheten av utvärderingarna.
- **Mer Detaljerad Utvärdering av LLM-baserade System:** Genomföra en djupare och mer specifik analys av LLM-baserade transkriptionsmodeller, inklusive deras beteende i realtid och deras förmåga att hantera specifika utmaningar som talarigenkänning (diarization) och hantering av överlappande tal. Detta skulle innebära att utveckla anpassade utvärderingsmetoder för denna nya typ av system.
- **Utveckling av Efterbearbetningspipelines:** Forskning och utveckling av robusta efterbearbetningsalgoritmer som automatiskt kan korrigera vanliga fel, förbättra meningssegmentering och eliminera "hallucineringar" i transkriptioner. Detta kan inkludera användning av språkmodeller för felkorrigering och segmentering av tystnader.
- **Studie av Finjustering för Små Språk:** En mer ingående studie av hur effektivt open-source-modeller kan finjusteras med mindre, specifika dataset för att förbättra prestandan för nischade domäner eller specifika accentgrupper inom svenska.
- **Kostnads- och Skalbarhetsanalys:** En djupgående analys av kostnaderna och skalbarheten för att implementera och underhålla både open-source och kommersiella **STT**-lösningar i en produktionsmiljö.

Referenser (Källförteckning)

Referenser till de använda modellerna och datakällorna kommer att inkluderas.

- https://huggingface.co/KBLab/kb-whisper-large
- https://elevenlabs.io/
- https://deepgram.com/
- https://azure.microsoft.com/
- https://huggingface.co/spaces/openai/whisper