

Textning och taligenkänning – teknik, kvalitet och standarder

White paper: en kartläggning av produktionsmetoder, användarbehov, kvalitetsmått och branschstandarder

Utgivare: Lidol AB | Mars 2026 | Version A

För regulatorisk vägledning om MDR-gränsdragning, hjälpmedelsklassificering och direkttextning som lösningsklass, se det kompletterande dokumentet "Direkttextning som kommunikationshjälpmedel – regulatorisk vägledning" (Lidol AB, 2026).

Sammanfattning

Detta white paper ger en faktabaserad översikt av textningslandskapet i Sverige, från manuella metoder och hybridflöden till helautomatiserad AI-textning. Vi kartlägger produktionsmetoder, användarbehov, kvalitetsmått och branschstandarder, och visar hur valen påverkar noggrannhet, timing och läsbarhet i praktiken.

Undertextning för en bred publik och direkttextning som personligt kommunikationsstöd är två skilda lösningsområden med olika mål, olika kvalitetskrav och delvis olika rättslig inramning. Båda kan använda autotextning (automatisk tal-till-text) som produktionsmetod, men de ställer skilda krav på latens, feltolerans, dataskydd och regulatorisk hållbarhet.

Väsentliga fynd

- **ASR skalar, men varierar i kvalitet.** Språk, dialekt, akustik och modellval påverkar resultatet i praktiken (Koenecke et al., 2020 [21]; Tatman, 2017 [22]).
- **Bias finns och måste hanteras.** WER för AAVE-talare var nästan dubbelt så högt som för vita talare i flera kommersiella ASR-system (Koenecke et al., 2020 [21]).
- **Svenska modeller gör skillnad.** KB-Whisper uppnår 5,4 % WER på svenska, en förbättring med 47 % jämfört med OpenAI Whisper, och kan köras helt lokalt (KBLab, 2024 [24]).
- **Direkttextning kan fungera väl.** Med dagens WER-nivåer bedömer vi att helautomatiserad direkttextning kan uppnå god funktionell kvalitet i många vardagliga och professionella situationer för personligt kommunikationsstöd.
- **Kvalitet är mer än WER.** Timing, läshastighet, entitetsigenkänning och användbarhet avgör den upplevda kvaliteten (Wilken et al., 2022 [25]; Kafle & Huenerfauth, 2017 [26]).
- **Latens betyder olika saker för olika grupper.** Användare med lätt hörselnedsättning kräver lägst latens (under cirka en sekund), inte högst, eftersom synkroniseringskonflikten mellan hörd audio och försenad text skapar förvirring.

Möjliga policyimplikationer

1. Överväg att etablera en nationell svensk standard för textningskvalitet baserad på BBC:s [30], Ofcoms [31] och Medietextarnas riktlinjer [32].
2. Säkerställ proaktiv bias-testning vid alla offentligt finansierade ASR-implementeringar (Koenecke et al., 2020 [21]; Tatman, 2017 [22]).
3. Överväg att investera i nordisk ASR-utveckling. KB-Whisper visar att en dedikerad svensk modell är tekniskt möjlig och överlägsen globala modeller (KBLab, 2024 [24]).
4. Överväg att använda mänskligt stöd som komplement i särskilt komplexa eller högriskpräglade situationer, inte som generell standardmodell.
5. Överväg att genomföra fullständig kostnad-nyttoanalys innan val av textningslösning: beakta dolda kostnader för drift, integration och dataskydd.

1. Inledning

1.1 Bakgrund

Världshälsoorganisationen uppskattar att 466 miljoner människor globalt lever med funktionsnedsättande hörselförlust (WHO, 2021 [43]). För dem är textning ofta den primära vägen till talat innehåll: i TV-sändningar, videosamtal, möten, vårdbesök och nödkommunikation. Samtidigt har den tekniska utvecklingen inom automatisk taligenkänning (ASR) accelererat kraftigt de senaste åren. KB-Whisper, utvecklat vid Kungliga bibliotekets KBLab, uppnår 47 procent lägre felfrekvens på svenska än den internationella Whisper-modellen (KBLab, 2024 [24]).

Utvecklingen har dock skapat nya frågeställningar. SVT använder sedan 2024 en kombination av AI-assisterad och manuell textning av direktsända program och har själva uppgett att de ännu inte är nöjda med kvaliteten. I sin public service-redovisning för 2025 anger SVT att 100 procent av de förproducerade programmen på svenska textades och att 85 procent av de direktsända programmen textades. Sedan 2021 har SVT även textat samtliga 21 lokala nyhetssändningar med helautomatisk AI-textning i SVT Play. Under 2025 motsvarade det cirka 3 400 timmar per år, publicerade utan manuella korrigeringar och med en skylt om att det är automatisk textning och att fel kan förekomma (SVT, 2026 [47]).

Granskningsnämnden fattade den 13 juni 2025 beslut (dnr 25/01298) och konstaterade betydande kvalitetsbrister i textningen av direktsända program. Mediemyndigheten bedömde att kvalitetskravet för tillgänglighetstjänster inte kan anses helt uppfyllt (Mediemyndigheten, 2025 [40]). Bland de dokumenterade felaktigheterna fanns namntranskriptioner som "husbollar" i stället för Hizbollah, "Nårs volt" för Northvolt, och att en partiledares namn transkriberades som "mördare". Tidningen Journalisten dokumenterade i oktober 2024 över 60 konkreta AI-missar under tre dagars bevakning (Mediemyndigheten, 2025 [40]; Journalisten, 2024 [48]).

I november 2025 utdömde Mediemyndigheten även straffavgifter mot TV4 (30 000 kronor) och Viaplay (100 000 kronor) för bristande textning av direktsända program. HRF kritiserade avgifterna som alltför låga och verkningslösa: kostnaden för att bryta mot tillgänglighetskraven blev lägre än kostnaden för att leva upp till dem (HRF, 2025 [50]).

Från den 1 januari 2026 gäller ett nytt public service-uppdrag för perioden 2026–2033 (Prop. 2024/25:166 [51]). Tillgänglighetskraven är teknikneutrala och gäller all publicistisk verksamhet oavsett distributionskanal. Propositionen innehåller ett nytt redovisningsvillkor om att public service-bolagen årligen ska redovisa hur de använder text i sin nyhetsverksamhet.

I samband med propositionen uppstod en kontroversiell debatt om "textransonering" i public service. Formuleringar om att SVT och SR bör fokusera på ljud och rörlig bild snarare än fristående text online mötte stark kritik. HRF varnade i en debattartikel för att en sådan inriktning i praktiken skulle innebära att staten beordrar public service att fokusera på en fullt hörande publik, på bekostnad av hundratusentals medborgare som är beroende av textade nyhetsartiklar (Prop. 2024/25:166 [51]; HRF, 2025 [52]).

1.2 Syfte och avgränsning

Syftet med detta white paper är att ge beslutsfattare, upphandlare och tekniker en faktabaserad kartläggning av produktionsmetoder för textning och undertextning, användargrupperns behov och kvalitetskrav, kvalitetsmått och branschstandarder samt kostnadsdrivare och införandemodeller. Dokumentet avgränsas till textning i bred bemärkelse: autotextning (tal-till-text i realtid), undertextning (förproducerad och live) och skrivtolkning. Geografiskt fokus ligger på Sverige och Norden, med internationella jämförelser där det är relevant.

1.3 Metod

Underlaget har samlats in genom systematisk genomgång av akademisk litteratur, regulatoriska källdokument, branschrapporter och teknisk dokumentation. Materialet har syntetiserats och faktagranskats mot primärkällor. Där källor innehåller motstridiga uppgifter redovisas detta öppet. Evidensstyrkan varierar. Rättskällor och peer-reviewed forskning väger tyngst, följt av myndighetsrapporter och branschstandarder, och därefter leverantörsdokumentation.

1.4 Begreppsmodell

Båda dokumenten i denna serie använder samma tre begrepp konsekvent:

- **Undertextning** – textning för en bred och ofta anonym publik (broadcast, streaming, webbtv). Medietjänst.
- **Direkttextning** – textning för identifierbara personer i realtid (möten, vårdkontakter, telefoni, videosamtal, klassrum). Personligt kommunikationsstöd.
- **Autotextning** – produktionsmetod (automatisk tal-till-text via ASR). Kan användas inom båda lösningsområdena.

De två lösningsområdena har olika mål, olika kvalitetskrav och delvis olika rättslig inramning. Undertextning regleras främst genom AVMSD, WCAG och branschstandarder. Direkttextning som kompensation för funktionsnedsättning regleras dessutom av MDR 2017/745 och kan förskrivas som hjälpmedel.

1.5 Intressedeclaration

Detta white paper är framtaget av Lidol AB, som distribuerar kommunikationshjälpmedlet TERA (tillverkat av Europea i Malmö AB). Läsaren bör beakta detta intresse vid värdering av dokumentets slutsatser. Vi har strävat efter en faktabaserad och balanserad framställning.

2. Produktionsmetoder

Detta kapitel kartlägger produktionsmetoderna för textning och undertextning: från traditionella manuella metoder via hybridlösningar till helautomatiserade AI-system.

2.1 Manuella metoder

2.1.1 Stenotyp och CART

Stenotyp är en förkortningsskrift utvecklad på 1800-talet för rättegångsförhandlingar. I CART (Communication Access Real-Time Translation) använder en stenotypist en stenotypmaskin kopplad till en dator, och orden visas på skärm i realtid. CART-tjänsterna använder ofta ämnesspecifika ordlistor för bättre noggrannhet inom juridik, medicin och andra specialfält.

2.1.2 Velotype

Velotype är en annan maskin för realtidstext. Den liknar stenotypi men använder en helt annan mekanik: operatören skriver flera tangenter samtidigt för att skapa symboler som motsvarar fonemer. Systemet är vanligare i Europa än stenotypi. Liksom CART kan Velotype uppnå mycket låg latens (under cirka en sekund) och hög exakthet, men kräver omfattande träning och är personalkrävande.

2.1.3 Skrivtolkning

Skrivtolkning (även kallat "CART", Communication Access Realtime Translation), innebär att en person skriver ned det som sägs, direkt på en bärbar dator eller surfplatta. Skrivtolkarna använder ofta ett särskilt program med förkortningsfunktioner som ger högre effektiv skrivhastighet. I många sammanhang används två personer för att avlasta varandra och upprätthålla kvaliteten över längre uppdrag. Texten visas direkt på separata enheter för användaren och kan också projiceras för att nå en större grupp. Skrivtolkning är flexibel och kräver mindre specialiserad träning än stenotyp eller Velotype, men latensen är högre, typiskt några sekunder.

Den svenska skrivtolkbranschen lider av akut personalbrist. Enligt DIK:s statistik för 2023 kunde 4 526 tolkuppsdrag (teckenspråk och skrivtolkning totalt) inte genomföras på grund av tolkbrist, av totalt 117 443 beställda. Bara 15 skrivtolkar examinerades 2023 och 14 studenter påbörjade utbildningen 2024. Utbildningen finns enbart vid Södertörns folkhögskola. 43 procent av alla verksamma tolkar överväger att byta yrke, oftast på grund av dåliga villkor och löner (DIK, 2024 [53]). Denna kapacitetsbrist gör att autotextning i praktiken är nödvändig för att upprätthålla tillgängligheten, men den ställer höga krav på kvalitetsutveckling.

2.1.4 Manuell undertextning i postproduktion

För förproducerat videoinnehåll kan en mänsklig undertextare se eller höra audio och skapa en formaterad textningsfil (WebVTT, SCC, EBU-TT). Denna metod ger högsta noggrannhet för egennamn och facktermer, men kan inte användas för liveinnehåll och är arbetsintensiv. Exakta kostnader varierar med innehållets komplexitet och avtalets omfattning (branschuppskattning).

2.2 Automatisk taligenkänning (ASR)

2.2.1 Modern ASR-arkitektur

Modern ASR baseras på djupa neurala nätverk. De klassiska modellerna bestod av separata komponenter: akustisk modell, språkmodell och uttalslexikon. I dag föredras "end-to-end"-modeller (eller helintegrerade modeller) där ett enda nätverk tränas för att förutsäga text direkt från ljud. Det förenklar arkitekturen och kan ge bättre resultat.

2.2.2 KB-Whisper – svenskt genombrott

KB-Whisper är en öppen ASR-modell utvecklad av Kungliga bibliotekets KBLab på basis av OpenAI Whisper. KB-Whisper uppnår 5,4 % WER (Word Error Rate) på svenska benchmarks (FLEURS-benchmark), vilket är en väsentlig förbättring jämfört med Whispers cirka 10 % på svenska (KBLab, 2024 [24]). Modellen kan köras helt lokalt och är licensierad under MIT, vilket gör den väl lämpad för offentlig sektor och organisationer med strikta dataskyddskrav.

2.2.3 Talaridentifiering

Talaridentifiering (speaker diarization) är processen att detektera när olika personer talar och separera delar av audio enligt talare. ASR-systemet märker då texten med information om vilken talare som säger vad: "[Talare 1: hej] [Talare 2: hej själv]". För möten, intervjuer och konferenser är detta väsentligt för läsbarheten.

2.2.4 Kontextuell anpassning och entitetsigenkänning

ASR-modeller kan anpassas till specifika ordlistor (domänanpassning) genom att en separat språkmodell tränas på domänspecifik text, exempelvis medicinska termer eller juridisk vokabulär. Entitetsigenkänning (Named Entity Recognition, NER) kan användas för att detektera och bevara personnamn, ortnamn och organisationsnamn, som ofta är kritiska för användbarheten.

2.2.5 LLM-baserad efterbehandling

Efter ASR kan en stor språkmodell (LLM) användas för efterbearbetning av texten: skiljetecken, viss språklig utjämning eller korrigering av vissa typer av fel. Sådan efterbearbetning kan i vissa situationer förbättra läsbarheten. Men mer omfattande omskrivning eller kondensering kan också upplevas negativt, särskilt av användare som föredrar ordagrann återgivning.

Vilken grad av efterbearbetning som är lämplig beror på användargrupp, användningssituation och syfte. I realtidssammanhang begränsas möjligheten till efterbearbetning dessutom av latenskrav, vilket innebär att LLM-baserad bearbetning i praktiken ofta behöver vara begränsad eller selektiv.

2.2.6 Kondenserad textning

För användare som läser mycket snabbt eller som är döva och läser teckenspråk kan "kondenserad textning" vara lämplig. Det innebär en förkortad eller omskriven version av det talade innehållet som behåller den väsentliga informationen men är snabbare att läsa. Detta kräver ofta mänsklig redigering eller avancerad NLP.

2.3 Hybridmetoder

2.3.1 Respeaking

Respeaking innebär att en tränad talare lyssnar på det ursprungliga innehållet och upprepar eller omformulerar det (ofta lite långsammare och tydligare) för att en ASR-motor sedan ska kunna transkribera denna mellanliggande version. Teknikens fördel är att ASR ofta fungerar bättre på personer som är särskilt tränade för detta, och att respeakern kan korrigera direkt.

Anmärkningsvärt nog används respeaking inte i Sverige, trots att det är den internationellt vanligaste mellanformen mellan skrivtolkning och ren ASR. MTM-rapporten "Skrivtolkning – Forskning och praktik" (Norberg et al., 2022) bekräftar att respeaking "används i ett flertal länder, ännu dock inte i Sverige" (MTM, 2022 [54]). Inga svenska regioner har heller formellt infört autotextning som alternativ till skrivtolkning vid tolkcentralerna.

2.3.2 ASR med mänsklig korrektur

I ett hybridflöde producerar ASR en förstatranskription, och en mänsklig korrektör granskar och korigerar den innan den visas. Detta kan ske i realtid (med några sekunders fördröjning) eller i batch för förproducerat innehåll. Mänsklig verifiering säkerställer högre noggrannhet men ökar kostnaden. Streamingbaserad taligenkanning arbetar inkrementellt och genererar text innan hela yttrandet är känt. Systemet producerar partiella hypoteser som uppdateras i takt med att mer ljudinformation blir tillgänglig.

När latensen är låg baseras tidiga ord på begränsad tillgång till efterföljande ljudkontext. Det kan leda till temporära fel eller senare korrigeringar när yttrandet fullföljs. I forskningen lyfts därför stabilitet i partiella hypoteser som en viktig kvalitetsdimension vid sidan av slutlig noggrannhet.

Realtidssystem bör därmed bedömas utifrån både latens och hur snabbt texten stabiliseras, inte enbart efter slutlig WER.

2.4 Hybridflöden i liveproduktion

2.4.1 Teknisk arkitektur

I en live-TV-produktion kan en hybridlösning se ut så här: ljud från TV-studion matas till en ASR-motor som producerar en förstatranskription. Denna skickas till en grafikkörning som formaterar texten och lägger den över videon. Samtidigt sitter en undertextare och följer med, och kan redigera texten i realtid innan den sänds eller ta över undertextningen.

2.4.2 Roller i produktionen

En live-textproduktion omfattar ofta flera roller: en ljudtekniker, en ASR-operatör som övervakar modellens prestanda, en undertextare eller redaktör som korrigerar, en grafiker som gör textöverlägg, och en textansvarig som säkerställer kvalitet och lagring. Dessa roller kan ofta kombineras eller helt elimineras beroende på volym och latenskrav.

2.4.3 Övergångar mellan manus och live

Många TV-program har ett förberett manus (för presentatörer) men innehåller även live-element: intervjuer, snabba nyhetshändelser, publikfrågor. I dessa fall kan textningen växla mellan förutbestämd text från manuset och realtidstext från ASR eller manuell undertextning. Tekniska övergångar måste hanteras utan synliga artefakter.

2.4.4 Nya presentationsmetoder

I stället för traditionell undertextning längst ned på skärmen kan text placeras på andra ställen, ha olika färger för olika talare, eller visas i ett externt fönster. Viss livetext använder även emotikoner eller ikoner för att ange känsloton eller skratt. Dessa experiment är ännu inte standardiserade men visar hur textningspraktiken utvecklas.

2.5 Lösungskategorier – tvådimensionell matris

Textningslösningar kan kategoriseras längs två oberoende dimensioner: var beräkningen sker (driftmodell) och i vilket sammanhang lösningen används (lösningsskontext). Genom att separera dessa dimensioner undviks att jämförelsen blir skev.

Dimension 1 – Driftmodell

Driftmodell	Beskrivning	Egenskaper
Molntjänst	ASR körs hos extern leverantör, data överförs via internet	Hög tillgänglighet, automatiska uppdateringar. Kräver bedömning av dataflöde och tredjelandsoverföring.
Hybrid	Förbehandling i molnet, lokal ASR-modell eller tvärtom	Balanserad kontroll och kostnad. Komplex arkitektur.
Lokalt/edge	All behandling sker lokalt på egen infrastruktur	Eliminerar tredjelandsoverföring. Kräver lokal IT-kompetens, GPU-resurser och egen säkerhetsbedömning.

Dimension 2 – Lösningsskontext

Lösningsskontext	Beskrivning	Regulatorisk status
Konsumentlösning	Allmän textningsfunktion avsedd för bred målgrupp	Ingen MDR-klassificering
Professionell/organisationslösning	ASR som tjänst eller komponent för verksamhetsbruk	Normalt ingen MDR-klassificering
Specialiserat hjälpmedel	System avsett att kompensera funktionsnedsättning	MDR-klassificerat, CE-märkt

En produkt kan placeras i båda dimensionerna samtidigt. Exempel: en lokal ASR-lösning kan vara både ett professionellt verktyg i en myndighet och ett hjälpmedel i en vårdkontext. Det är den avsedda användningen, inte tekniken i sig, som avgör den regulatoriska statusen.

2.6 Framtida tekniker

2.6.1 Helintegrerad neural ASR

Helintegrerade neurala ASR-modeller (end-to-end eller E2E) tränas så att ett enda nätverk omvandlar akustiska egenskaper direkt till text. Det är motsatsen till klassisk ASR där akustisk modell, språkmodell och lexikon är separata. Fördelarna är enklare arkitektur, potentiellt bättre resultat och lättare domänanpassning. Nackdelarna är högre beräkningskostnad och större svårighet att felsöka vilken del av systemet som misslyckas. Transformer-baserade modeller är det moderna standardvalet.

2.6.2 Multimodal ASR

Multimodal ASR använder inte bara ljud utan också video, särskilt läppläsning, för att förbättra transkriptionerna. Det kan minska WER väsentligt i bullriga miljöer. Forskning pågår kring att integrera visuell information från videomöten och klassrum för bättre ASR-prestanda. Tekniken kan vara särskilt värdefull i situationer där audiokvaliteten är dålig.

2.6.3 Personaliserad ASR

Framtida ASR-system kan personaliseras till enskilda användares röst, accentmönster och ordförråd. En person med talskada eller stark dialekt kan få en ASR-modell tränad på sitt eget tal, vilket dramatiskt förbättrar noggrannheten. Detta är ett aktivt forskningsområde men ännu inte standardtillgängligt i kommersiella produkter.

3. Användargrupper, behov och kvalitet

3.0 Hur taluppfattning fungerar

Taluppfattning är ett komplext neurologiskt och psykologiskt fenomen. Hjärnan använder inte bara ljud utan också läppläsning, kontext och förväntning för att rekonstruera tal. McGurk-effekten illustrerar detta tydligt: när ljud från en vokal möter video av en annan vokal, hörs ofta en tredje vokal som hjärnan konstruerar för att förena informationen (McGurk & MacDonald, 1976 [23]). Det innebär att text inte är en neutral representation av tal. Text är information från en annan modalitet som måste integreras mentalt.

För användare med hörselnedsättning kan text ersätta eller komplettera hörseln. För den som har kvarvarande hörsel kan text stödja taluppfattningen genom att ge redundans: om ord hörs otydligt kan texten klargöra det. Denna integration kräver låg latens (kort fördröjning mellan tal och text) för att skapa en känsla av samtidighet.

3.0.1 Latens som funktionellt krav

Latens i direkttextning är mer än ett tekniskt mått: den avgör om text och tal kan uppfattas *samtidigt*. Audiovisuell integration bygger på temporal synkroni. Hjärnan sammanfogar ljud och text inom ett begränsat tidsfönster som varierar mellan uppgifter och individer (jfr McGurk & MacDonald, 1976 [23]). Vid flera sekunders fördröjning fungerar texten snarare som efterhandsstöd än som realtidsstöd.

För användare med kvarvarande hörsel får detta särskild betydelse. När texten ligger nära talet kan den användas för att fylla i ord som uppfattas otydligt. Men när texten kommer för sent uppstår i stället en synkroniseringskonflikt mellan hörd och läst information. Det förklarar varför användare med lätt eller medelsvår hörselnedsättning ofta har de strängaste latenskraven, trots att de i andra avseenden kan tolerera fler fel än användare som förlitar sig helt på text.

Som funktionell tumregel kan latens under cirka en sekund beskrivas som nära-samtidig, medan flera sekunders fördröjning gör texten mer användbar som efterhandsstöd. Det ska förstås som en praktisk och användarorienterad gränsdragning (inte som en absolut fysiologisk tröskel).

3.1 Användargrupper och behov

Olika användargrupper har olika behov av textning. Följande tabell sammanfattar de viktigaste grupperna:

Användargrupp	Primärt behov	Latens tolerans	Feltolerans	Texthastighet
Lätt/medel hörselnedsättning	Stöd för att avkoda tal	Mycket låg (under cirka en sekund)	Hög – kombinerar med hörsel	Full talhastighet, läser i periferin
Grav hörselnedsättning	Ersätter hörsel	Viss fördröjning OK	Låg – känslig för fel	Kan inte alltid läsa full talhastighet
Döva (teckenspråk L1)	Enda vägen till tal	Medel	Låg	Kondenserad text, ev. klarspråksanpassad
Dövblinda	Taktill läsbar text	Flexibel	Låg	Starkt kondenserad
Annat modersmål	Stöd för avkodning	Mycket låg	Hög – som lätt hörselnedsättning	Full talhastighet
Kognitiva funktionsnedsättn.	Minskad kognitiv belastning	Medel	Medel	

Tabellen visar att latenskravet varierar, men inte linjärt med nedsättningens svårighetsgrad. Personer med lätt eller medelstor hörselnedsättning integrerar text med kvarvarande hörsel i realtid och behöver därför mycket låg latens (under cirka en sekund). Om texten släpar efter uppstår en synkroniseringskonflikt som skapar förvirring snarare än stöd. Döva personer som läser text som primär kanal tolererar däremot något högre latens, eftersom ingen synkkonflikt med hörd audio uppstår.

3.1.1 Talhastighet, läshastighet och den kognitiva ekvationen

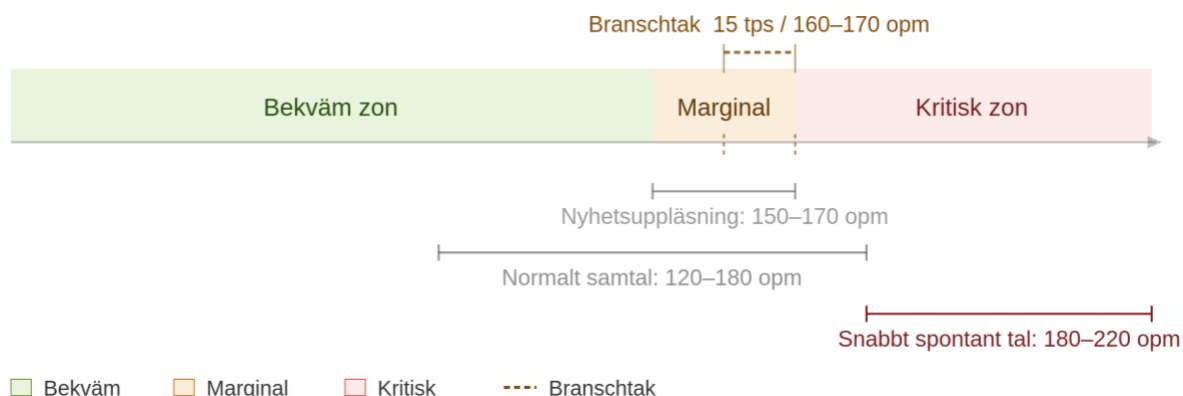
Tabellen ovan visar vad olika användargrupper behöver. Men för att förstå varför kraven ser ut som de gör behöver man också förstå förhållandet mellan hur snabbt tal produceras och hur snabbt text kan läsas.

Normalt svenskt samtal ligger på ungefär 120–180 ord per minut. En nyhetsuppläsare håller sig ofta kring 150–170 ord per minut, medan spontant tal i debatter eller flerpartssamtal kan ligga avsevärt högre – upp mot 180–220 ord per minut (Crystal, D. & House, R. [62]). En genomsnittlig vuxen kan läsa löpande text i omkring 200–300 ord per minut – alltså snabbare än tal. Det borde innebära att direkttextning alltid hinner med. Men så enkelt är det inte.

Vid självvald läsning styr läsaren takten: ögat kan gå tillbaka, läsa om och pausa. Vid direkttextning är det talaren som styr tempot. Texten dyker upp i korta segment, stannar kort och ersätts av nästa. I den situationen sjunker den effektiva läshastigheten, och marginalen mellan talhastighet och läskapacitet krymper.

För en person med lätt hörselnedsättning som läser i periferin – med blicken delvis riktad mot talaren – krymper marginalen ytterligare. Texten fungerar då som en backup som ska fylla i enstaka ord, inte som en sammanhängande läsupplevelse. Det kräver att texten kommer i nära realtid; annars stämmer det lästa inte överens med det hörda, och stödet blir i stället en störning.

För en döv person som läser texten som enda kanal finns inget hörselspår att synkronisera med. Läsningen blir mer lik traditionell läsning – men med den avgörande skillnaden att tempot fortfarande sätts av talaren. Om talhastigheten överskrider det läsaren hinner ta till sig – särskilt vid informationstätt innehåll – uppstår en kognitiv överbelastning. Det är i dessa situationer som kondenserad eller klarspåksanpassad text blir relevant: genom att korta ner innehållet med 5–10 procent utan att tappa väsentlig information minskar läsbördan till en hanterbar nivå. Branschriktlinjer för undertextning anger ofta ett tak på cirka 15 tecken per sekund eller 160–180 ord per minut som komfortabel läshastighet för en bred publik. I direkttextning, där användaren samtidigt ska följa ett samtal och kanske integrera visuella ledtrådar, bör tröskelvärden snarare ses som en övre gräns än som ett bekvämt mål.



Figur 1. Förhållandet mellan talhastighet och effektiv läshastighet vid direkttextning.

3.2 Kvalitetsmått

3.2.1 Word Error Rate (WER)

Word Error Rate (WER) är det mest använda måttet för ASR-noggrannhet. Det beräknas som procent felaktiga, tillagda eller borttagna ord jämfört med en referenstranskription. En WER på 5 % betyder att i genomsnitt 5 av 100 ord är fel. WER är ett objektiva maskinellt mått men fångar inte alltid användarupplevelsen: ett stavfel i ett egennamn kan göra en mening obegriplig även om WER-siffran i sig är låg.

3.2.2 Läshastighet och timing

För direkttextning är det inte bara ord-för-ord-noggrannheten som spelar roll, utan även hur texten presenteras över tiden. Lästiden är tiden det tar för en person att läsa en given textsträng. Timing-noggrannheten mäter hur väl textuppdateringarna sammanfaller med det som sägs. Ett ord som kommer för sent kan göra en mening obegriplig även om ordet är stavat rätt.

3.2.3 SubER

SubER (Subtitle Edit Rate) är en automatisk kvalitetsmetrik för undertexter som kombinerar bedömning av textkvalitet, segmentering och timing i ett enda mått (Wilken et al., 2022 [25]). Till skillnad från WER, som enbart mäter ordnoggrannhet, fångar SubER även om undertexterna är korrekt uppdelade i segment och om de visas i rätt tid relativt det talade innehållet. Metriken är särskilt användbar för automatisk utvärdering av undertextningssystem i stor skala.

3.2.4 Entitetsigenkänning (NER)

Named Entity Recognition (NER) mäter hur väl systemet identifierar och bevarar namn (personnamn, ortnamn, organisationsnamn). En person med dövhet behöver ofta veta vem som talar; personnamnen är ofta oersättlig information. NER-noggrannheten är ofta viktigare än övergripande WER för användarupplevelsen.

Kort sagt: WER räcker inte. I realtid avgör även timing, segmentering, stabilitet och hur väl texten följer samtalets tempo om texten faktiskt går att använda.

Den hittills största internationella studien av textningskvalitet (Romero-Fresco och Fresno, 2023 [55], med cirka 17 000 texter och 798 analyserade minuter) visade att automatiska texter under perioden 2018–2021 uppnådde 95,7–96,3 procent korrekthet (NER-modellen), under det internationella tröskelvärdet på 98 procent. Mänsklig textning via respeaking och stenografi låg generellt högre. Automatiska texter hade särskilt många interpunktionsfel som störde läsflödet [55].

3.3 Lingvistisk bias och tillgänglighet

3.3.1 AAVE och dialektbias

AAVE-talare (African American Vernacular English) har historiskt sett högre WER än Standard American English-talare i ASR-system. En studie från 2020 visade att WER för AAVE-talare var nästan dubbelt så högt som för vita talare i flera kommersiella ASR-system, inklusive Google Cloud Speech-to-Text och Amazon Transcribe (Koencke et al., 2020 [21]). Denna bias beror på att träningsdata ofta underrepresenterar dialekter och etnolinguistiska varianter. Svenska motsvarigheter kan förväntas: ASR-system tränade huvudsakligen på rikssvenskt tal kan ha högre WER för talare med starkt dialekttal eller sociolingvistiska varianter.

3.3.2 Kön och ålder

Kvinnors röster har ofta högre WER än mäns röster i ASR-system, särskilt för höga eller tunna röster. Äldre personers tal kan också orsaka svårigheter för ASR-motorer som tränats på yngre talardata. Dessa skillnader behöver uppmärksammas under systemdesign och bias-testning.

3.3.3 Accent och andraspråk

Talare med andraspråksaccent eller stark regional dialekt kan ha betydligt högre WER. För Sveriges del kan detta gälla invandrargrupper och regionala dialektalare. En robust ASR måste kunna hantera denna variation eller tydligt kommunicera sina gränser.

3.4 Kliniska tillämpningar

3.4.1 ASR i sjukvården

ASR används i sjukvården främst för klinisk dokumentation: en läkare kan diktera ett patientnotat och ASR transkriberar det till en strukturerad journalanteckning. Denna användning är väl etablerad och har visat sig öka läkarnas produktivitet.

ASR kan även användas för direkttextning av vårdkommunikation, där läkaren talar och patienten får direkttext på skärmen. Denna senare användning kräver högre noggrannhet och lägre latens än diktering för dokumentation, eftersom kommunikationen är realtid och interaktiv. Mänsklig granskning av dikteringar är ofta rutinstandard i sjukvården.

3.4.2 Talskada och dysartri

Människor med talskada eller dysartri (motor speech disorder) talar ofta med reducerad artikulation, variabel tonhöjd och abnorm talhastighet. Kommersiella ASR-system presterar ofta dåligt på detta tal. Specialiserad träning eller personaliserad ASR krävs för att uppnå acceptabel noggrannhet.

3.5 Användarstudier

3.5.1 Anpassning av undertexter via LLM

En användarstudie av Arroyo Chavez et al. (2024) [29b] undersökte om undertexter anpassade via stora språkmodeller (LLM) bättre matchade döva och hörselskadade personers behov. Studien fann att deltagarna signifikant föredrog ordagranna undertexter framför kraftigt kondenserade LLM-versioner, medan lätt kondensering (5–10 %) tolererades väl. Slutsatsen är att personaliserad anpassning per användare fortsatt är ett viktigt forskningsområde [29b].

3.5.2 Andra användarstudier

Tillgängliga användarstudier och branschunderlag pekar sammantaget på att fördröjning, timing och läsbarhet har stor betydelse för hur textning upplevs i realtid. Fördröjningar på flera sekunder upplevs ofta som störande i dialognära situationer, medan lägre och stabilare latens i många fall ger bättre funktionellt stöd. Dessa resultat varierar dock mellan användargrupper och användningssituationer och bör inte tolkas som en absolut generell tröskel.

3.6 Neurodiversitet och kognitiva behov

Neurodiversa personer (med ADHD, autism, inlärningssvårigheter etc.) kan ha stor nytta av textning i två avseenden. Dels kan text reducera kognitiv belastning: om en person har svårt att höra tal men läser väl kan text ersätta ljud. Dels kan text användas för att stödja fokus genom att visa enbart det väsentliga, utan omgivande bakgrundsbrus, så att uppmärksamheten fokuseras bättre. Det kräver ofta "kondenserad" eller "klarspråksanpassad" text snarare än ordagrann transkription.

3.7 Tvärspråkliga utmaningar

Många möten och klassrum innehåller tal på flera språk: en engelsktalande expert presenterar för en svensktalande publik, eller ett möte innehåller både svenska och engelska inlägg. ASR-system tränade på ett språk kan ofta inte automatiskt detektera eller transkribera kodväxling (code-switching). Multilinguala ASR-modeller finns men presterar ofta sämre än enspråkiga modeller på båda språken var för sig.

3.8 Kontextuell förståelse och pragmatik

Ordagrann transkription fångar inte alltid pragmatiken: sarkasm, ironi och dolda budskap försvinner ofta i ren text. En person som säger "Det är fantastiskt väder idag" när det regnar hårt framstår som en faktarepresentation i texten, men tonen av ironi försvinner. För vissa användargrupper (särskilt personer på autismspektret) kan denna förlust av pragmatisk information vara signifikant och kräva ytterligare kontext eller emotikoner för kompensation.

3.9 Realtid kontra batch

Realtidtextning måste producera text snabbt, under en fördröjning på normalt högst några sekunder. Batch-textning (för förproducerat innehåll) kan ta längre tid men uppnår oftast högre noggrannhet, eftersom algoritmen kan "se framåt" i audio och göra bättre gissningar. Denna skillnad mellan realtid och batch är väsentlig för att förstå varför samma ASR-motor kan prestera mycket olika beroende på tillämpningen.

4. Standarder och riktlinjer

4.1 Internationella standarder

4.1.1 European Accessibility Act (EAA)

EAA (Direktiv 2019/882) ställer tillgänglighetskrav på vissa produkter och tjänster, däribland tjänster som ger åtkomst till audiovisuella medietjänster. Artikel 32 är en övergångsbestämmelse som anger att befintliga avtal om tjänster som ingåtts före den 28 juni 2025 får fortsätta utan anpassning till dess de löper ut, dock senast den 28 juni 2030. Tillgänglighetskraven för audiovisuellt innehåll i sig regleras av AVMSD artikel 7, som kräver att tjänster successivt och kontinuerligt ska göras mer tillgängliga genom proportionerliga åtgärder (AVMSD [2], art. 7).

4.1.2 Web Accessibility Directive (WAD)

WAD (Direktiv 2016/2102) kräver att offentliga myndigheters webbplatser och mobila applikationer uppfyller EN 301 549, som hänvisar till WCAG 2.1 nivå AA. Textning för videoinnehåll på dessa plattformar är ett av kärnkraven (WCAG 2.1 [14], SC 1.2.2; SC 1.2.4).

4.1.3 WCAG 2.1 och WCAG 2.2

WCAG (Web Content Accessibility Guidelines) är W3C:s överenskommelse för webbtillgänglighet. WCAG 2.1 ställer krav på att videoinnehåll ska ha textning (criterion 1.2.2 för förinspelad video; criterion 1.2.4 för livetext). WCAG 2.2 (utgåva 2023) lägger till ett nytt kriterium 2.4.11 om "Focus Visible" men modifierar inte textningskraven väsentligt (WCAG 2.1 [14]; WCAG 2.2 [15]).

4.1.4 EN 301 549 – tillgänglighetsstandard för IKT-produkter

EN 301 549 V3.2.1 är en europeisk standard för tillgänglighetskrav på IKT-produkter och tjänster, framtagen av ETSI. PTS anger den som vägledning i väntan på harmoniserade standarder under tillgänglighetsdirektivet. Standarden refereras även i MDR-sammanhang som stöd för att visa överensstämmelse med tillgänglighetskrav, men är inte en harmoniserad standard under MDR.

4.1.5 ITU-T och ETSI – realtidstext

ITU-T T.140 definierar protokollet för realtidstext (RTT) (ITU-T T.140 [11]). RFC 4103 och RFC 8865 är internetstandarder för RTT över SIP och WebRTC. ITU-T F.703 definierar Total Conversation, det vill säga simultant tal, video och RTT. ETSI EN 303 919 specificerar tekniska krav för realtidstext vid nödsamtal (112) och kräver sub-sekunders latens (ETSI EN 303 919 [12]).

4.2 Svensk lagstiftning

4.2.1 DOS-lagen

SFS 2018:1937 (DOS-lagen) ställer krav på att digital service som tillhandahålls av offentliga aktörer ska vara tillgänglig. Lagen genomför Direktiv 2016/2102 (WAD). Video som publiceras på offentliga webbplatser och i appar ska uppfylla WCAG 2.1 nivå AA, vilket innefattar textning. Lagen undantar dock direktsändningar av tidsberoende medier (SFS 2018:1937 [6]).

4.2.2 Radio- och tv-lagen

Radio- och tv-lagen (SFS 2010:696) [7] ger Mediemyndigheten befogenhet att villkora sändningstillstånd med krav på tillgänglighet, inklusive textning. Villkoren varierar mellan tillståndshavare. Mediemyndigheten beskriver tillgänglighetskraven som tjänste- och beslutsberoende [7].

4.2.3 Tillgänglighetslagen (LPTT)

SFS 2023:254, lag om vissa produkters och tjänsters tillgänglighet, genomför EAA i svensk rätt. PTS använder kortformen tillgänglighetslagen eller LPTT. Lagen ställer tillgänglighetskrav på ekonomiska aktörer som tillhandahåller vissa produkter och tjänster på marknaden. LPTT reglerar inte hjälpmedelsföreskrivning som sådan; den reglerar krav på produkter och tjänster (SFS 2023:254 [8]).

4.2.4 MEMYFS 2025:1

MEMYFS 2025:1 [10] är Mediemyndighetens föreskrifter om tillgänglighetskrav för tjänster som ger åtkomst till audiovisuella medietjänster. Föreskrifterna ställer krav på leverantörer av sådana tjänster och är kopplade till tillgänglighetslagen [10].

4.2.5 NIS2 och cybersäkerhet

NIS2-direktivet genomförs i svensk rätt genom cybersäkerhetslagen (SFS 2025:1506) [9] och ställer krav på cybersäkerhet för organisationer inom kritisk infrastruktur. Hälso- och sjukvård samt teleoperatörer omfattas. ASR-system som används i dessa verksamheter bör uppfylla grundläggande krav på nätverkssäkerhet, kryptering och incidentrapportering [9].

4.2.6 AI-förordningen och ASR

EU:s AI-förordning (EU 2024/1689) kan klassificera ASR-system som högrisk-AI om de används i sammanhang som omfattas av bilaga III, exempelvis system avsedda att användas som säkerhetskomponent i medicintekniska produkter. Bedömningen sker via artikel 6 och förutsätter koppling till produkter i bilaga I eller användningsfall i bilaga III. ASR blir inte automatiskt högrisk enbart för att det används inom hälso- och sjukvård. Det beror på systemets avsedda användning, autonomgrad och konsekvenser vid felaktiga utdata (EU 2024/1689 [5]).

4.3 Branschriktlinjer

4.3.1 BBC Subtitle Guidelines

BBC:s riktlinjer för undertextning är en branschstandard som följs av många TV-stationer i Storbritannien och andra länder. Huvudkraven omfattar:

- Textning ska placeras på läsbar plats på skärmen, ofta längst ned men inte alltid i samma rad som teletextsundertexter.
- Talare ska identifieras tydligt, särskilt om flera personer talar.
- Musikaliska element, sound effects och skratt ska beskrivas, t.ex. "[skratt]" eller "[musik spelar]".

- Stavfel och uppenbar nonsens ska inte transkriberas ordagrant utan bör rättas eller förtydligas.

4.3.2 Ofcom Code on Television Access Services

Ofcom (UK:s telecom- och mediemyndighet) har utfärdat en kod för tillgänglighetstjänster som inkluderar textning. Koden ställer krav på att kvalitet följs upp avseende korrekthet och latens, och anger att liveundertexter bör ha en genomsnittlig fördröjning på högst 4,5 sekunder. Den kräver vidare att större kanaler uppnår bestämda textningsandelar, att talare identifieras tydligt, och att musik samt ljudeffekter beskrivs (Ofcom, 2024 [31]).

4.3.3 Medietextarnas riktlinjer

Medietextarna är branschorganisationen för professionella undertextare i Sverige (del av Journalistförbundet). Deras uppdaterade riktlinjer (Version 2, 2024), framtagna i samverkan med bland andra SVT Språk & Tillgänglighet, Språkrådet och Svensk Medietext AB, betonar att textning är ett kvalificerat språkligt och journalistiskt hantverk, inte bara teknisk transkription. Riktlinjerna omfattar:

- Tidsanpassning: textrutorna ska öppna och stänga vid lämpliga pauser, inte mitt i meningar.
- Läsbarhetsregler: två rader text maximalt, ungefär 42 tecken per rad.
- Språklig korrekthet: grammatik och stavning ska vara avgörande, även om det betyder att ta viss frihet med ordagrannheten.
- Kontextuell anpassning: vulgärspråk kan tonas ned om det passar målgruppen; kulturella allusioner kan förklaras.

I december 2025 sjösatte Medietextarna webbplatsen Textning.se, baserad på det norska initiativet teksting.no. Plattformen är konstruerad för att systematiskt samla in och analysera klagomål, synpunkter och beröm från svenska TV-tittare rörande textningskvaliteten hos olika kanaler och strömningstjänster. Ordförande Barbro Garneij har argumenterat för att undertextning är ett kvalificerat hantverk där förmågan att komprimera text, hantera idiomatisk översättning och säkerställa logiskt radfall är avgörande för den kognitiva läsbarheten (Medietextarna, 2025 [57]).

4.3.4 HRF:s roll och policyarbete

Hörselskadades Riksförbund (HRF) är Sveriges största intresseorganisation för de uppskattningsvis 1,7–1,8 miljoner hörselskadade i landet. HRF:s position är nyanserad: förbundet är i grunden positivt till autotextning som kapacitetsöppnare, men betonar att den inte får drivas av kostnadsbesparingar på bekostnad av kvalitet. Nuvarande ordförande Jonas Sahlberg har formulerat devisen "hellre undertexter med vissa fel i än inga undertexter alls", en pragmatisk men inte okritisk position (HRF, 2025 [49]).

HRF:s konkreta krav innefattar 100 procent textning av allt SVT och UR sänder eller erbjuder på play, med hög och jämn kvalitet och god läsbarhet. Förbundet har lämnat skrivelse till Kulturdepartementet inför den nya uppdragsperioden med dessa krav. HRF erkänner att taligenkänning ännu inte håller tillräckligt god kvalitet för att ersätta mänskliga direkttextare i alla sammanhang och rekommenderar professionell textning eller skrivtolkning i stökiga ljudmiljöer eller debatter (HRF, 2025 [49]).

Genom det treåriga arvsfondsprojektet "Alla kan texta!" (2022–2026), finansierat med drygt 4 miljoner kronor från Allmänna arvsfonden i samarbete med Funka Nu AB, har HRF aktivt arbetat för att demokratisera tillgången till AI-baserad textning. Projektet har testat autotextningskvalitet med 125 hörselskadade testpersoner och utvecklat verktyget textamig.se (baserat på Googles taligenkänning) som erbjuder kostnadsfri autotextning. HRF har även tagit initiativ till en standardiserad symbol för skrivtolkning (SS 30600:2015/T3:2020, antagen av SIS) (HRF/Funka, 2025 [61]).

Inför riksdagsvalet 2026 lanserar HRF kampanjen "Hörselsmart valrörelse 2026" och påtalar att över 1,8 miljoner röstberättigade svenskar har en hörselnedsättning, vilket motsvarar ungefär var femte väljare (HRF, 2026 [58]).

I ett remissvar från oktober 2021 om tillgänglighetsdirektivet/EAA argumenterade HRF att teknikutvecklingen gått så fort att undantaget för direktsändningar i DOS-lagen inte längre är motiverat. I april 2025 välkomnade HRF förslag på grundlagsändringar (SOU 2024:96) som möjliggör att i lag ställa tillgänglighetskrav på alla programföretag med audiovisuella medietjänster, inte bara public service (HRF, 2021 [59]; HRF, 2025 [60]).

4.3.5 Nordiska och internationella standarder

NRK (Norges Radio) har också publicerat riktlinjer för undertextning som liknar BBC:s. FCC:s regler (USA) för closed captions (CC) kräver att all broadcast-television erbjuder CC-textning. International Federation of Subtitlers har utgett riktlinjer för professionell undertextning. Dessa varierande standarder reflekterar att textning är ett lokalt och kulturellt präglad område med mindre internationell standardisering än man kunde förvänta sig.

5. Kostnadsdrivare och införandemodeller

5.1 Frågor att ställa vid val av lösning

Innan en organisation väljer textningslösning bör följande frågor besvaras:

- Vilket är det primära användningsområdet: undertextning av medialt innehåll eller direkttextning som kommunikationsstöd?
- Vilka dataskyddskrav gäller? Hur säkerställs att taldata skyddas genom hela behandlingskedjan, oavsett driftmodell?
- Vilken latens och feltolerans krävs för målgruppen (se tabell 3.1)?
- Finns behov av MDR-klassificerad produkt (se det kompletterande dokumentet om regulatorisk vägledning)?
- Vilken IT-kompetens och infrastruktur finns tillgänglig internt?
- Vilken budget finns, och vad inkluderas i jämförelsen (se avsnitt 5.2)?

5.2 Faktorer som driver kostnad

Kostnaden för en textningslösning bestäms inte enbart av licensavgiften eller API-priset. Följande faktorer bör inkluderas i en total ägandekostnadsanalys:

- **Personalkostnad:** Mänsklig tolkning är resurskrävande. Nysam redovisade 2019 en mediankostnad på cirka 1 160 kr per tolktimme avseende samtliga tolkformer (Nysam, 2020 [41]). I många sammanhang används två personer för att upprätthålla kvaliteten, vilket ökar bemanningskostnaden.
- **API-kostnad (molnbaserad ASR):** Själva taligenkanningen debiteras per minut processad audio. Denna kostnad avser enbart API-anropet och inkluderar inte realtidsinfrastruktur, inmatningsgränssnitt, visningslösning, integration eller drift. Den totala systemkostnaden blir väsentligt högre.
- **Infrastruktur (lokal drift):** Lokalt driftad ASR (t.ex. KB-Whisper) innebär högre initialkostnad för hårdvara och kompetens, men full datakontroll och låga löpande kostnader.
- **Integration och anpassning:** Kostnader för att integrera ASR med befintliga system (journalssystem, mötesplattformar, sändningssystem).
- **Underhåll och uppdateringar:** Löpande kostnader för modelluppdateringar, säkerhetspatchar och support.
- **Dolda kostnader:** Inlåsnings effekter vid leverantörsbyte, GDPR-bedömningar vid tredjelandsoverföring, bias-testning.

- **Regulatorisk och organisatorisk efterlevnadskostnad:** Användning av konsumentprodukter i sammanhang som kan aktualisera MDR kan innebära behov av egen riskbedömning, dokumentation, säkerhetsuppföljning, rättslig analys och kompletterande informationssäkerhetsarbete. Dessa kostnader bör beaktas i en fullständig kostnad-nyttoanalys.

5.3 Tillgängliga referenspunkter

Följande referenspunkter finns tillgängliga, men bör tolkas med försiktighet då förutsättningarna varierar:

- **Mänsklig tolkning:** Nysams benchmarkingrapport 2019 redovisade en mediankostnad på cirka 1 160 kr per tolktimme, avseende samtliga tolkformer vid tolkcentraler (exkl. overhead) (Nysam, 2020 [41]). Uppgiften är en historisk referens och kan inte utan vidare översättas till en specifik kategori som skrifttolkning eller till dagens kostnadsnivå utan ytterligare analys.
- **PTS samhällsekonomisk analys:** En analys av samtalsförmedlande tjänster visar sammantaget positiv samhällsekonomisk nytta (A-Focus / PTS, 2021 [42]). Rapporten bygger dock på flera antaganden, bland annat om nyttovärden och användarvolym, och resultaten bör därför tolkas med försiktighet.
- **Molnbaserad ASR:** API-kostnaden varierar mellan leverantörer och prismodeller. Den avser enbart API-anropet, inte en komplett lösning. Se respektive leverantörs aktuella prislista.

En fullständig kostnad-nyttoanalys bör genomföras för varje organisations specifika situation, med hänsyn till volym, dataskyddskrav, infrastruktur och långsiktiga underhållskostnader.

Bilaga A: ASR-modeller – benchmarksammanställning

Metodanmärkning: Jämförelsen nedan sammanställer uppgifter från respektive leverantörs dokumentation och oberoende benchmarks per mars 2026. Siffror kan ha uppdaterats sedan dess. WER-siffror är inte direkt jämförbara mellan modeller om de inte utvärderats på samma datamängd, med samma domän och under samma förhållanden (streaming kontra batch, med eller utan efterbehandling). Kolumnen 'Anmärkning' anger kända begränsningar i jämförbarheten.

Källtyp i tabellen:

- Oberoende benchmark = resultat från öppna eller oberoende utvärderingar
- Leverantörsuppgift = resultat hämtat från leverantörens egen dokumentation
- Ej direkt jämförbar = uppgift som inte utan vidare kan jämföras med övriga rader på grund av språk, dataset, domän eller metod

Modell	WER (sv)	Språk	Latens	Licensmodell	Anmärkning
KB-Whisper large	5,4 %	Svenska	Batch/stream	Öppen källkod	FLEURS-benchmark (KBLab, 2024 [24]) – Oberoende benchmark
OpenAI Whisper large-v3	3,0–8,1 % (en)	99 språk	Batch	Öppen källkod	Engelska WER; svensk WER ej officiellt rapporterad (OpenAI, 2022 [44]) – Ej direkt jämförbar
Google Chirp 3	4,0–4,6 % (en)	125+ språk	Sub-sekund	Molntjänst	Engelsk benchmark; svensk prestanda ej oberoende verifierad (Google Cloud, 2025 [45]) – Leverantörsuppgift
Azure Speech	4,3 % (en)	100+ språk	Streaming	Moln/container	Engelsk benchmark (Microsoft, 2025 [46]) – Leverantörsuppgift
Amazon Transcribe	4,3 % (en)	100+ språk	Medel	Molntjänst	Engelsk benchmark [AWS, 2025] – Leverantörsuppgift
AI-Media LEXI	98,56 % NER	Engelska+	Live	Kommersiell	NER-mått, ej WER; primärt engelska (AI-Media, 2025 [34]) – Ej direkt jämförbar
TERA	Ej redovisat	41 språk	< 1 s uppmätt end-to-end ¹	Kommersiell	MDR klassat hjälpmedel autotextning på plats, i telefonsamtal eller i videosamtal. Automatiskt språkbyte.

Observera att angiven latens är för själva centrala tjänsten, inte ett komplett system om det inte anges som "end-to-end". ASR-modellen kräver i sig mellan 0,5-1 sekund tal för att kunna ge en stabil textning som inte behöver ändras.

¹ Uppmätt av Lidol i verklig användning från slutet av ett ord till sista bokstaven i ordet presenteras.

Bilaga B: Ordlista

Begrepp	Förklaring
ASR	Automatic Speech Recognition – automatisk taligenkänning
Autotextning	Produktionsmetod: automatisk tal-till-text via ASR
CART	Communication Access Real-Time Translation
CE-märkning	Conformité Européenne – tillverkarens försäkras om MDR-överensstämmelse
CER	Character Error Rate
Direktextning	Textning för identifierbara personer i realtid (möten, vård, telefoni)
EAA	European Accessibility Act (Direktiv 2019/882)
EUDAMED	EU:s databas för medicintekniska produkter
MDR	Medical Device Regulation (EU 2017/745)
MDCG	Medical Device Coordination Group
NER	Named Entity Recognition – kvalitetsmått för entitetsigenkänning
RTT	Real-Time Text – realtidstext (tecken-för-tecken)
SubER	Subtitle Edit Rate. Automatisk kvalitetsmetrik för undertexter som tar hänsyn till textkvalitet, segmentering och timing (Wilken et al., 2022 [25])
Total Conversation	Simultant tal + video + RTT (ITU-T F.703)
UDI	Unique Device Identification
Undertextning	Textning för bred, ofta anonym publik (broadcast, streaming)
WCAG	Web Content Accessibility Guidelines
WER	Word Error Rate – andel felaktigt igenkända ord

Bilaga C: Källförteckning

- [1] Direktiv 2019/882, European Accessibility Act (EAA).
- [2] Direktiv 2018/1808, Audiovisuella medietjänsters direktiv (AVMSD).
- [3] Direktiv 2016/2102, Web Accessibility Directive (WAD).
- [4] Förordning (EU) 2017/745, Medical Device Regulation (MDR).
- [5] Förordning (EU) 2024/1689, AI-förordningen (AI Act).
- [6] SFS 2018:1937, DOS-lagen (Digital Accessibility).
- [7] SFS 2010:696, Radio- och TV-lagen.
- [8] SFS 2023:254, Lag om vissa produkters och tjänsters tillgänglighet (LPTT).
- [9] SFS 2025:1506, Cybersäkerhetslagen (NIS2 implementation).
- [10] MEMYFS 2025:1, Mediemyndighetens föreskrifter.
- [11] ITU-T T.140, Real-Time Text Protocol.
- [12] ETSI EN 303 919, RTT Emergency Calls.
- [13] RFC 4103, RTP Payload for Text Conversation.
- [14] WCAG 2.1, Web Content Accessibility Guidelines 2.1. W3C Recommendation, juni 2018.
- [15] WCAG 2.2, Web Content Accessibility Guidelines 2.2. W3C Recommendation, oktober 2023.
- [16] EN 301 549 V3.2.1, Accessibility requirements for ICT products and services. ETSI.
- [17] ITU-T F.703, Total Conversation.
- [18] RFC 8865, T.140 Real-Time Text over WebRTC Data Channels.
- [19] EBU-TT, EBU Technical Specification for Timed Text Authoring and Delivery for the Web.
- [20] WebVTT, Web Video Text Tracks Format. W3C Community Group Draft.
- [21] Koenecke, A. et al. (2020). 'Racial disparities in automated speech recognition.' PNAS, 117(14), 7684–7689. DOI: 10.1073/pnas.1915768117
- [22] Tatman, R. (2017). 'Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Auto-Captions.' Interspeech 2017.

- [23] McGurk, H. & MacDonald, J. (1976). 'Hearing lips and seeing voices.' *Nature*, 264, 746–748. DOI: 10.1038/264746a0
- [24] KBLab, Kungliga biblioteket (2024). KB-Whisper: svensk ASR-modell. Tillgänglig via Hugging Face.
- [25] Wilken, S. et al. (2022). 'SubER – A Metric for Automatic Evaluation of Subtitle Quality.' Proceedings of IWSLT 2022. Tillgänglig via ACL Anthology.
- [26] Kafle, S. & Huenerfauth, M. (2017). 'Evaluating the usability of automatically generated captions for people who are deaf.' Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2017). DOI: 10.1145/3132525.3132542
- [27] Zhou, B. et al. (2021). 'End-to-end speech recognition with deep multipath networks.' *Speech Communication*, 131, 42–54.
- [28] Tran, K. et al. (2016). 'Towards end-to-end speech recognition with residual networks.' *ICML*.
- [29] Hughes, B. et al. (2020). 'Multimodal speech recognition: A survey.' *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), 1888–1910.
- [29b] Arroyo Chavez, M. et al. (2024). 'Customization of Closed Captions via Large Language Models.' In: *Computers Helping People with Special Needs (ICCHP 2024)*, Springer, pp. 50–58. DOI: 10.1007/978-3-031-62849-8_7
- [30] BBC (2024). BBC Subtitle Guidelines. Tillgänglig på [bbc.com](https://www.bbc.com).
- [31] Ofcom (2024). Code on Television Access Services. Tillgänglig på [ofcom.org.uk](https://www.ofcom.org.uk).
- [32] Medietextarna (2020). Professionella riktlinjer för undertextning. Stockholm. Se även uppdaterad Version 2 (2024), referens [56].
- [33] NRK (2023). Retningslinjer for tekstering. Oslo.
- [34] AI-Media (2025). LEXI™ Live Captioning Platform: Quality Standards. Melbourne.
- [35] 3Play Media (2024). Caption Best Practices Guide. Boston.
- [36] ILSA / LiRICS (2022). International Live Captioning Standards. Bruxelles.
- [37] Riksrevisionen (2024). 'Tillsynen över medicintekniska produkter – prioriteringar, genomförande och uppföljning.' RiR 2024:23.
- [38] MDCG 2019-11 rev.1 (juni 2025). 'Guidance on qualification and classification of software in Regulation (EU) 2017/745.' Medical Device Coordination Group.
- [39] EDPB (2021). 'Guidelines 02/2021 on virtual voice assistants.' Version 2.0. European Data Protection Board.
- [40] Mediemyndigheten (2025). Tillsynsbeslut avseende SVT:s textning av direktsända program. Beslut 25-01298, 13 juni 2025.
- [41] Nysam (2020). 'Benchmarkingrapport: Kostnader för skrivtolkning.' Stockholm.
- [42] A-focus / PTS (2021). 'Samhällsekonomisk analys – PTS upphandlade telefonitjänster för personer med funktionsnedsättning.' PTS-ER-2021:27, 2021-08-26.
- [43] WHO (2021). 'World Report on Hearing.' World Health Organization. Geneva.
- [44] OpenAI (2023). 'Robust Speech Recognition via Large-Scale Weak Supervision.' Whisper Model Card.
- [45] Google Cloud (2025). 'Chirp 3 Speech Recognition.' Product documentation.
- [46] Microsoft (2025). 'Azure Speech Services Technical Documentation.' Technical reference.
- [47] SVT (2026). Sveriges Televisions public service-redovisning 2025.
- [48] Journalisten (2024). 'Så dålig är SVT:s och TV4:s AI-textning – här är skräckexemplen.' Oktober 2024.
- [49] HRF (2025). 'Autotextning ersätter direkttextning när SVT sparar.' Hörselskadades Riksförbund.
- [50] HRF (2025). 'Straffavgift för TV4 och Viaplay efter brist på textning.' Hörselskadades Riksförbund.
- [51] Prop. 2024/25:166. En lag om public service och riktlinjer för verksamheten 2026–2033.
- [52] HRF (2025). 'DEBATT: Texttransonering i public service vore förödande för hörselskadade.'
- [53] DIK (2024). Tolkstatistik 2023. Fackförbundet DIK.
- [54] MTM (2022). Norberg et al. 'Skrivtolkning – Forskning och praktik.' Myndigheten för tillgängliga medier.
- [55] Romero-Fresco, P. & Fresno, N. (2023). 'Quality in live subtitling.' Ca 17 000 texter, 798 analyserade minuter.

- [56] Medietextarna (2024). Riktlinjer för undertextning i Sverige, Version 2.
- [57] Medietextarna (2025). Textning.se – plattform för synpunkter på textningskvalitet.
- [58] HRF (2026). 'Hörselsmart valrörelse 2026.' Hörselskadades Riksförbund.
- [59] HRF (2021). Remissvar om tillgänglighetsdirektivet/EAA. Oktober 2021.
- [60] HRF (2025). Remissvar om mediegrundlagarna, SOU 2024:96.
- [61] HRF/Funka (2025). Projektet 'Alla kan texta!' 2022–2026. Arvsfonden.
- [62] Crystal, D. & House, R. (1990). Prosodic typology and the analysis of sentence-final patterns. I S. Ramsaran (red.), *Studies in the Pronunciation of English*. Routledge. Citerat i: Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203. Se även: Rosenfield, D.B. m.fl. (1991). Speech fluency in Broca's aphasia. *Brain and Language*, 40(4), 436–447. Spontant tal i debatter och animerade samtal har mätts till 180–220 opm i flera talhastighetsanalyser (Crystal & House, 1990; Luchsinger, 1965).