Niclas Ringbom

40605

Tillämpning av automatisk röstigenkänning för nordiska språk

**Abstrakt**

I dagens värld har automatisk röstigenkänning förankrat sig starkt i vad vi ser som nutida och framtida teknologi, i och med den kraftigt ökande populariteten för användningen av virtuella assistenter. Eftersom automatisk röstigenkänning länge var främst riktad mot engelskspråkiga målgrupper är syftet med denna avhandling att, på basis av tidigare undersökningar, recensera för pålitligheten hos nordiska ASR-system. I denna avhandling beskrivs även strukturen för ett typiskt ASR-system och alla dess huvudkomponenter. Populära verktyg och dataset för utveckling introduceras kort.

[**1. Introduktion**](#_vux1mvczy7nt) **4**

[**2. Bakgrund**](#_kqj43i3mbp8q) **5**

[2.1 Historia](#_ae31ke9q1qm2) 6

[2.2. Grundläggande strukturen för ASR](#_apkfzzb00qzo) 7

[2.3. Särdragsextrahering](#_gx1o1jgqw9fu) 8

[2.4. Akustisk modellering](#_8l157iwzef04) 9

[2.5. Språkmodellen](#_2ptvm4v78lzq) 10

[**3. Metoder för ASR**](#_a0imvy2hbw11) **11**

[3.1 Gaussian Mixture Model – Hidden Markov Model GMM-HMM](#_prgujpnmwah0) 11

[3.2 Deep Neural Network DNN](#_bhci7plj94jw) 11

[3.3 End-to-end ASR](#_u9kq4k8givi6) 11

[**4. Open source hjälpmedel**](#_nkthc2j7e5q8) **11**

[4.1 Open source verktyg](#_ozop2m8zjkm1) 12

[4.1.1 Kaldi](#_4fqfwesji89j) 12

[4.1.2 Espresso](#_cl85nsnp1in4) 12

[4.2 Open source träningsdata](#_7chuu7thner0) 13

[4.2.1. LibriSpeech](#_v6cgiphttt3c) 13

[4.2.1. WSJ Wall Street Journal](#_h4753hm11g4f) 13

[4.2.3. NST (Språkbankens resurskatalog)](#_40tss42aurry) 14

[**5. Tillämpningar av ASR**](#_s0stjq450qpw) **15**

[5.1 Engelskspråkiga applikationer](#_vnq55u93i7yk) 16

[5.2 Nordiska applikationer](#_qvgdyx448ivq) 17

[**6. Utveckling av ASR för nordiska språken**](#_71ms9jio131k) **18**

[6.1 Tidigare undersökning](#_4lfi3nhhblyo) 18

[6.1.1 Svensk undersökning](#_7qcixk3po472) 18

[6.1.2 Finsk undersökning](#_3tyrwnymy4hy) 19

[6.2 Förhinder](#_i7qyqqe3d6o7) 19

[6.3 Förbättringskrav](#_5lleamft72ua) 20

[**7. Diskussion**](#_74fxrj56ytqj) **20**

# 1. Introduktion

Automatisk röstigenkänning (ASR) har som följd av behovet att effektivisera och underlätta vardagliga uppgifter, återtagit en central roll i mjukvaruproduktion. I stället för att välja eller skriva in kommandon för ett program kan konsumenten använda sin röst för att styra och beordra programmet. Detta är en stor fördel i situationer där konsumenten inte har fria händer eller för personer med begränsad rörlighet. Att tillämpa automatisk röstigenkänning för en uppgift krävs endast någon sorts interaktion mellan en dator och en människa. I dagens värld finns det otaliga tillämpningar och ännu fler möjligheter som följd av teknologins centrala roll i vårt dagliga liv.

Många storspelare inom teknologibranschen utnyttjar automatisk röstigenkänning, oftast i samband med en artificiell intelligens. Av dessa är de mest kända Apples Siri, Microsofts Cortana och Amazons Alexa. Dessa tre använder sig av automatisk röstigenkänning på liknande sätt, och försöker få kommunikationen mellan människa och maskin att vara så naturlig som möjligt. Kommunikationen påbörjas med ett nyckelord som väcker upp apparaten och dess röstigenkänningsmodul. Detta följs av ett kommando, fråga eller dylikt, exempelvis "Alexa, play music".

Inom automatisk röstigenkänning tillämpas många olika metodologier och teknologier för att skapa en produkt som bäst passar till en specifik uppgift. De främsta faktorerna som påverkar valet av metod och algoritm är vare sig modellen önskas optimeras för ett fåtal personer eller flera, hur stort ordförråd röstigenkänningsmodulen bör behärska och hur rent det mottagna talet är, det vill säga vare sig det finns bakgrundsbrus eller inte. Dessa problem studeras närmare i kapitel 2.3

Utöver de kännetecknande faktorerna som definierar funktionaliteten hos ett ASR-program existerar ett större hinder. Detta huvudsakliga hinder för automatisk röstigenkänning är behovet för en enorm mängd data för att träna modellen att förstå sig på språket, kontexten och uttalen. För engelskan är detta inte ett problem eftersom det finns otaliga mängder data, men för mindre språk är det svårare att hitta tillräckligt med data av hög kvalité att träna en djupinlärningsmodell med.

Med dessa hinder och framgångsrika tillämpningar i åtanke är således syftet med denna avhandling att undersöka och redogöra för tillämpningar av automatisk röstigenkänning för nordiska språk samt jämföra dessa med engelskspråkiga varianter. Problem och hinder karaktäristiska för ASR hos nordiska språk introduceras även i denna avhandling.

# 2. Bakgrund

Inom detta kapitel beskrivs och förklaras den grundläggande strukturen för röstigenkänning, indelad i den akustiska modelleringen, språkmodelleringen samt särdragsextraheringen. En kort uppläggning av historian för automatisk röstigenkänning ges även för att ge läsaren en inblick i hur mycket teknologin har utvecklats.

Inom ASR existerar det otaliga typer av röstigenkänning; nyckelordsigenkänning, talarberoende och oberoende system etc. Det önskade ordförrådet påverkar även valet av typ stort eftersom WER(Word Error Rate) ökar ifall systemet inte är begränsat inom ett visst område (definieras ofta som context dependent och indepedent) [4].

Vid uvecklingen av ett ASR system måste en huvudsaklig typ väljas på basis av otaliga faktorer, och detta val måste oftast definieras om flera gånger under utvecklingsprocessen för att få en så passande struktur som möjligt. Ett bra exempel är testen utförda av N. Vanhainen och G. Salvi för deras akustiska modell [4]. Ur Figur 2.0 kan det observeras att data med lägre samplingsfrekvens gav den akustiska modellen sämre precision i form av WER. En annan viktig observation är den stora skillnaden mellan kontextfria modeller och kontextberoende. På basis av liknande test kan utvecklaren välja rätt data och metod för sin modell.

Figur 2.0 Sammanfattning av testresultat för olika data, samplingsfrekvenser och metoder [4].

## 2.1 Historia

Syftet med detta kapitel är att bekanta läsaren med historian för automatisk röstigenkänning för att lättare förstå hur långt ämnet har utvecklats under det senaste decenniet. En bredare uppfattning om tidigare tekniker underlättar även inlärningen av de nyare, mera komplicerade teknikerna samt hur dessa tekniker tillämpas i dagens teknologi.

I och med att teknologin utvecklades klarade röstigenkänningssystemen av att behandla större ordförråd samt mera krävande sorter av tal, bland annat anslutna siffror och kontinuerligt tal. När utveckling av ASR-system påbörjades på 1960-talet kunde endast enstaka isolerade ord kännas igen [6]. De första teknikerna som användes var Filter-bank analysis, Time-Normalization och Dynamic Programming men redan på 1970-talet började man använda sig av Pattern Recognition som DNN (Deep Neural Network) fortfarande tillämpar i dagens läge. På 1980-talet introducerades n-gram språkmodeller som var och är en ytterligt viktig komponent i ASR-system.

Uppkomsten av Hidden Markov modeller på 1980-talet tillät systemen att uvecklas tillräckligt långt för att kunna processera kontinuerligt tal. Från och med 90-talet blev användningen av ändliga automater och statistisk inlärning, vilka vidare utvecklades till maskininlärning i början av 2000-talet.

Inom det senaste årtiondet har trenden gått mot användning av GMM-HMM, DNN och end-to-end ASR system. Dessa tekniker introduceras i kapitel 3, Metoder för ASR.

## 2.2. Grundläggande strukturen för ASR

I dagens värld existerar många olika metoder och variationer av ASR men de bygger alla på samma basmetodologi. Grunden består av två huvudsakliga delar, särdragsextrahering och klassificering [1], där särdragsextraheringens uppgift är att dela upp talet i rimliga delar, oftast korta ljudsnitt som beskriver en enstaka bokstav eller siffra. Klassificeringens uppgift är sedan på basis av ljudsnittet bestämma vilken bokstav, siffra eller ord ljudsnittet med störst sannolikhet korrelerar med.



Figur 1. Arkitekturen hos ett typiskt röstigenkänningssystem

I figur 1 beskrivs den så kallade "pipelinen" eller arkitekturen hos ett röstigenkänningssystem. Processen påbörjas av ett intag av data från användaren i form av röstsignaler som sedan skickas vidare till särdragsextraheringen. Särdragsextaheringen delar upp datan och förbereder den för klassificeringsmodulen. Inom klassificeringsmodulen är första steget att tolka de uppdelade fönstren, detta utförs steg för steg. Tolkningen använder sig av tidigare information för att koppla ihop stycken(fönster) till ord. Efter denna sammankoppling producerar den akustiska modellen och språkmodellen en gemensam hypotes som definierar det mest sannolika ordet **w\*** [3]**.**

## 2.3. Särdragsextrahering

Särdragsextraheringen är ett extremt viktigt steg i röstigenkänningssystem, eftersom data som fås in i systemet oftast varierar stort. Olika användare, bakgrundsbrus, varierande hårdvara och akustik är några av faktorerna som påverkar datan som ges in till systemet och alla dessa variabler påverkar igenkänningen stort. Det är således viktigt för särdragsextraheringen att samla in och stycka upp datan på ett sådant sätt att klassificeringen kan arbeta med liknande data, oberoende av dessa försvårande variabler.

Som följd av att särdragsextraheringen spelar en så central roll har det utvecklats många olika metoder varav de främsta är följande: Linear Predictive Analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive Coefficients (PLP), Mel-Frequency Cepstral Coefficients (MFCC), Power spectral analysis (FFT), Mel scale cepstral analysis (MEL), Relative spectra filtering of log domain coefficients (RASTA), first order derivative (DELTA) [1]. I denna avhandling sätts fokus på MFCC varianten eftersom det är en populär metod hos röstigenkänningssystem tillämpade för svenska [3][4][5].

Mel Frequency Cepstral Coefficients (MFCC) presenterades på 1970-talet av Mermelstein [7], men idén för dessa koefficienter existerade redan före detta. Syftet med MFCC är att dela upp inmatningen i 25-30 ms fönster [8]. Detta skapar ett korttids talspektrum vars fonem kan vidare behandlas. Utvecklingen av MFCC var starkt bunden till frekvensanalysen för taluppfattningen hos människoörat, som lättare uppfattar skillnader i ton på lägre frekvenser, vilket ledde till att det beslöts att använda mera filtrering för spektrum med låga frekvenser och mindre filtrering på de högre frekvenserna. Den så kallade mel skalan reflekterar detta, vilken är linjär på låga frekvenser och logaritmisk på höga frekvenser.

En variant av MFCC utvecklades även från samma bas, där logaritmen av filtrens resultat behandlades, och inte själva resultaten som MFCC använder. Denna variant kallas "log mel-frequency spectral Coefficients"eller MFSC [8]. MFCC särdrag har visat sig vara speciellt lämpliga för HMM-GMM ASR-system eftersom särdragen är dekorrelerade, som GMM med diagonal kovarians föredrar.

## 2.4. Akustisk modellering

Den akustiska modellen representerar (i par med språkmodellen) ett ASR-systems hjärna. Det är här den egentliga tolkningen av uttal sker och detta lyckas genom att träna modellen med talinspelningar samt transkripten för uttalen. Träningen av den akustiska modellen är ett av hindren när ASR-system tillämpas för språk med färre tillgängliga resurser. För att träna en bra akustisk modell krävs hundratals gigabyte av ren, transkriberad data med tillräckligt hög samplingshastighet och resolution. Oftast används 16kHz och 16 bit för dessa två eftersom det är gynnbart att träna modellen på samma samplingshastighet och resolution som systemet förväntas tolka. Dessa värden används eftersom de lätt kan uppnås utan att kräva professionell studioutrustning (främst ljudkort och mikrofon).

Efter att särdragsextraheringen har behandlat inmatningsdatan är det den akustiska modellens uppgift att tolka uttalen för varje fönster och ge ut en rekommendation för vilket fonem uttalet mest sannolikt korrelerar med. Eftersom alla ord byggs upp av under hundra grundläggande fonem (antalet varierar beroende på språket) kräver estimering av fonem betydligt mycket mindre beräkning än hela ord, även om det i vissa fall kan försvåra igenkänningen ifall talaren har en dialekt eller accent obekant för den akustiska modellen. Detta orsakar problem för små språk med många olika dialekter där det inte finns tillräckligt med data för att träna den akustiska modellen på.

Ett alternativ till träning på enstaka fonem (monofonem) är trifonem där trifonemet är en sekvens på tre fonem som tillåter modellen att lättare uppskatta fonemet i mitten av kedjan, vilket ger bättre träffsäkerhet. Trifonem är generellt sett som ett bättre alternativ än monofonem och används oftare i utvecklingen av akustiska modeller [3][9].

## 2.5. Språkmodellen

Efter att uttalet har tolkats och konverterats till fonem med hjälp av särdragsextraheringen och den akustiska modellen är det språkmodellens uppgift att estimera sannolikheten för sekvensen av ord [5]. Språkmodellen estimerar med andra ord sannolikheten att ordet "hungrig" följer frasen "Jag är" med hjälp av tidigare träningsdata. Detta underlättar igenkänningen eftersom det utesluter ord från tolkningsberäkningarna, exempelvis att en mening med låg sannolikhet slutar på ordet "och". Med hjälp av språkmodellen får ASR-systemet en uppfattning av kontext, beroende på metoderna använda och hur bra modellen är tränad.

Likt de andra delarna i ett ASR-system existerar det många olika slags språkmodeller, varav n-gram språkmodellen är en av de mest populära och mest undersökta metoderna för röstigenkänning med stor vokabulär [3][5]. Eftersom ett stort vokabulär är en klar fördel för röstigenkänningsmjukvara behandlas främst n-gram språkmodellen i denna avhandling.

N-gram modellen använder sig av tidigare n antal ord för att bestämma sannolikheterna för nästa ord [3]. Modellen lär sig genom att analysera ordsekvenserna i träningsdatan, vilket leder till att en stor varians i meningsstruktur och längd är till stor nytta. Användning av orensad eller ensidig träningsdata leder till att språkmodellen gör antaganden om det uttalade språket som inte reflekterar språkets egentliga natur. Samling av perfekt träningsdata är tyvärr omöjligt och för att förhindra n-gram med orealistiskt hög sannolikhet använder man sig av så kallade "smoothing techniques" [5]. Dessa motverkar att modellen lär sig förvänta specifika ord, och på så vis förvränger meningen eller ordet.

# 3. Metoder för ASR

Efter introduktionen av de grundläggande delarna för ett ASR-system som behandlades i kapitel 2 är det dags att gå in djupare på de olika metoderna som kan användas för att förverkliga ett fungerande system. De tre populäraste metoderna är i dagens läge GMM-HMM, DNN och end-to-end varianterna. Detta kapitel börjar med att introducera den mest populära och historiskt betydande metoden GMM-HMM och beskriver sedan DNN och end-to.end metoderna, som har haft stor framgång under det senaste årtiondet.

## 3.1 Gaussian Mixture Model – Hidden Markov Model GMM-HMM

WIP

## 3.2 Deep Neural Network DNN

**WIP**

## 3.3 End-to-end ASR

# 4. Open source hjälpmedel

Eftersom det krävs relativt mycket resurser för att producera ett fungerande ASR-system, kan det vara till stor hjälp att inte behöva samla all data själv, välja rätt särdragsextraheringsmetod och träna modellerna. För att underlätta detta har många forskare inom ASR och NLP gett ut sina lösningar i form av öppen källkod. Dataset med öppen licens är även till stor fördel vid utveckling av ASR-mjukvara. Detta kapitel introducerar två populära verktyg, Kaldi och Espresso, samt beskriver några instanser av träningsdata med öppen licens.

## 4.1 Open source verktyg

### 4.1.1 Kaldi

Ett av de mest populära open source ASR-verktygen är Kaldi [20], som är skrivet i C++ och är licensierat under Apache License v2.0. Kaldi är utvecklad främst för forskare i NLP och ASR, vilket innebär att verktyget är lätt att modifiera för ett specifikt forskningssyfte. Detta innebär även tyvärr att Kaldi är ett svårt verktyg att använda, och kräver utförlig erfarenhet i ASR-system och deras utveckling. Kaldis utvecklare upprätthåller detaljerad dokumentation och har en aktiv utvecklar bas på Github [21]. Tidigare forskning utförda av Emilie Kullmann [3] och Zimon Mossberg [5] ansåg att utvecklarna bakom Kaldi är responsiva och gärna svarade på frågor.

Kaldi använder sig av traditionella metoder för att uppnå röstigenkänningen, exempelvis användningen av GMM-HMM eller DNN för att träna modellerna. Kaldi använder sig av så kallade "recipes" för att definiera olika variationer av modellerna. Dessa används för att definiera Kaldi instanser som stöder specifika språk, exempelvis Emelie Kullmans recept för svenska finns tillgänglig på [22]. På så vis kan andra utvecklare förbättra tidigare recept eller modifiera dem för att tolka ett nytt språk.

### 4.1.2 Espresso

Espresso [23] är ett end-to-end ASR verktyg som lanserades 2019 av Y. Wang et.al. [24]. Verktyget baserar sig på det populära PyTorch biblioteket och Fairseq, ett neuralt maskinöversättningsverktyg [24]. Espresso uppnår högklassiga resultat på de populära dataseten WSJ, LibriSpeech och Switchboard, utan att modifiera datan [24].

Fördelen med Espresso är att det är byggt på ett enda djupinlärningsramverk (PyTorch) och är således lätt att utvidga på eller integrera i applikationer. Utöver detta erbjuder Espresso möjligheten att utföra distribuerad inlärning på flera GPUn (Graphics Processing Unit).

## 4.2 Open source träningsdata

En väsentlig del av hela utvecklingsprocessen är tillgängligheten till stora mängder röstdata samt deras sammankopplade transkript. Denna data krävs för att utveckla själva igenkänningsmodulen och används även för att testa träffsäkerheten för den producerade mjukvaran. I detta kapitel introduceras allmänt accepterade dataset för engelskan samt några nordiska varianter.

### 4.2.1. LibriSpeech

LibriSpeech är en av de mest använda dataseten för ett engelskspråkigt ASR-system, ihopsamlad av V. Panayotov et. al. [26]. LibriSpeech korpuset gavs ut till allmänheten 2015 och består främst av data från LibriVox projektet. Datan extraherades från ljudböcker och datasetet innehåller ungefär 1000 timmar av tal med 16 kHz samplingsfrekvens[26].

LibriSpeech är starkt kopplad till Kaldi verktyget tack vare Povey, som medverkar i både LibriSpeech och Kaldi. I introduktionsartikeln för datasetet [26] visar de att akustiska modeller tränade på LibriSpeech datan producerar en mera träffsäker modell än modeller tränade på WSJ (Wall Street Journal) datasetet. Datasetet finns tillgängligt under CC BY 4.0 licensen på OpenSLRs resurskatalog[25].

### 4.2.1. WSJ Wall Street Journal

TODO

### 4.2.3. NST (Språkbankens resurskatalog)

Det norska nationalbiblioteket har under åren samlat in många dataset lämpliga för ASR-system, och erbjuds på norska, svenska och danska [27]. Dataseten finns tillgängliga i olika samplingsfrekvenser, varav 16 kHz är den mest använda. Datasetet för svenska består av ungefär 500 timmar av svensk tal [3] och inkluderar många olika dialekter.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Purpose** | **Lines** | **Persons** | **Recordings** | **Size(GB)** |
| Training | 312 | 920 | 287040 | 96,5 |
| Testing | 987 | 80 | 78960 | 22,7 |

Figur 3, NST Svenska Datasetets egenskaper

Röstdatan består av dikterat tal, inbandat i en studio med en mikrofon, vilket förminskar närvaron av yttre störningsfaktorer, som bakgrundsbrus eller flera talare samtidigt [5]. NSTs dataset är en av de få resurser tillgängliga för utveckling av ASR-system för de nordiska språken och är således aktivt använt inom området, exempelvis av Emelie Kullmann [3], Zimon Mossberg [5] och G. Salvi et. al. [4]. Deras resultat tas upp närmare i kapitel X, Tidigare undersökning.

# 5. Tillämpningar av ASR

Som det nämndes i avhandlingens introduktion finns det otaliga tillämpningar för röstigenkänningssystem, vilka kan delas in i olika former av transkribering, röstdrivna användargränssnitt samt identifieringsmetoder. Dessa olika typer av röst- och taligenkänning varierar stort från varandra, synnerligt i storleken på deras ordförråd. Som följd av att kraven på ordförråden är olika, använder de även olika varianter av röstigenkänning.

 Mjukvara fokuserad på att transkribera fritt tal kräver ett stort ordförråd, och kan beroende på applikationen tillämpa inlärning på användarens talkaraktärestik för att lättare uppfatta talet. Liknande komplexitet krävs inte för applikationer med specifika kommandon eftersom kravet på ordförråd är betydligt lägre. Nackdelen med dessa applikationer är att användaren förväntas memorera de flesta kommandon, vilket gör programmet ointuitivt. Detta, i kombination med lusten att få ASR-system att verka så "mänskliga" som möjligt har lett till att de ledande utvecklarna för applikationer med röstdrivna användargränssnitt i dagens läge implementerar fritt tal.

Utöver transkriberingen och de röstdrivna användargränssnitten tillämpas röstigenkänning, eller mera specifikt talarigekänning, för att identifiera en användare. För att förverkliga identifiering av en användare med hjälp av tal används samma tekniker som för röstigenkänning, men med större fokus på frekvenssvängningarna i talet. Uppbyggnaden av systemet kräver således inte nödvändigtvis en lika komplicerad systemarkitektur. Röstbiometrik, som området kallas, används ofta i multifaktor autentisering. Eftersom språket inte har stor inverkan på identifieringsprocessen behandlas endast applikationer som tillämpar antingen transkriberingsvarianten samt röstdrivna applikationer i denna avhandling.

I de följande kapitlen, 5.2 och 5.3, introduceras några applikationer för att ge läsaren en tydligare bild om hur ASR är tillämpat i dagens läge.

## 5.1 Engelskspråkiga applikationer

Nya utvecklingar inom teknologibranschen är oftast producerade med engelskspråkigt användargränssnitt, och ASR-programvara är inget undantag. Programvaran och modellerna testas ofta på engelska för många olika orsaker. På stora internationella företag är oftast redan engelskan kontorspråket och utvecklingen av programvara på detta språk underlättar således kommunikationen på arbetsplatsen. En annan fördel av att utveckla en ASR-applikation för engelska är den otroliga mängden data tillgänglig, både gratis och köpbara paket av data. En klar åtskiljning kan göras mellan mängden data tillgänglig på engelska och exempelvis svenska med att se på Mozillas "Common Voice" projekt där användare får fritt banda in och verifiera tal. I projektet finns för tillfället 38 GB data tillgängligt på engelska, medan det endast finns 153 MB av data på svenska [11].

Det finns även en klar ekonomisk fördel av att utveckla applikationer för engelska eftersom denna målgrupp är betydligt större och lättare att nå. I dagens internationella värld kan det nästan antas att alla kan engelska till någon mån, vissa nationaliteter bättre än andra. Här uppstår den grundläggande frågan och ett stort hinder för ASR-utveckling för mindre språk, varför spendera tid och resurser till utveckling av en ASR-applikation för ett mindre språk, t.ex. finska, när denna målgrupp delvis kan nås av en engelskspråkig applikation?

Marknaden för transkribering på engelska har stor konkurrens, där de största teknologiföretagen (Google, Microsoft etc.) tävlar mot varandra med sina egna speech-to-text mjukvaror. Även mindre kända företag har lyckats nå en betydlig andel av marknaden, exempelvis Dragon [13] och Speechmatics [14].

## 5.2 Nordiska applikationer

De största spelarna på ASR marknaden, Google, Microsoft, Apple, Amazon, har alla utvecklat programvara för engelska och är nu i processen att använda samma programvara för att tolka mindre språk. Googles speech-to-text platform täcker redan så gott som alla medelstora språk, inklusive alla större nordiska språk (svenska, finska, norska, danska, isländska) [12]. Microsofts språkplattform erbjuder, utöver speech-to-text stöd, färdiga språkmodeller för de nordiska språken. Apples Siri stöder även

Automatisk röstigenkänning för de nordiska språken har gjort stora framsteg under de senaste åren, utöver teknologijättarnas framgång har även företag baserade i Norden klarat av att producera konkurrenskraftig ASR-mjukvara för finska, svenska, norska och danska. För att vidare förbättra sin position inom NLP (Natural Language Processing) i Norden, har flera samarbetsprojekt påbörjats. I Sverige påbörjades ett sådant projekt November 2019 med målet att erbjuda verktyg för Sveriges myndigheter för integreringen av NLP i nuvarande och framtida tjänster [15]. Projektet har flera samarbetspartners, bland annat RISE [18] och Peltarion [19]. Projektet finansieras av Vinnova [16], Sveriges innovationsmyndighet, vilket reflekterar Sveriges engagemang för integreringen av NLP i det dagliga livet.

I Finland grundades FAIA (Finland's Artificial Intelligence Accelerator) av Näringsministeriet och Teknologiateollisuus. FAIAs syfte är att stöda företag i implementeringen av AI och NLP [17]. Dessa samarbetsprojekt visar att både Sverige och Finland är engagerade i att förbättra teknologin för ASR och NLP, samt är villiga att implementera detta i statliga tjänster.

# 6. Utveckling av ASR för nordiska språken

Syftet med detta kapitel är att närmare undersöka på vilken nivå ASR för de nordiska språken är för tillfället, samt redogöra för faktorer som begränsar framsteg. En recension för tidigare undersökningar presenteras för att ge konkreta exempel på resultat. I samband med överblicken på hindren ges även några förbättringskrav för att nordisk ASR ska nå en maximal nivå.

## 6.1 Tidigare undersökning

I detta kapitel recenseras några undersökningar gjorda för de nordiska språken, språkvis indelade.

### 6.1.1 Svensk undersökning

Svenskspråkig undersökning för ASR är starkt kopplad till dataseten tillgängliga på NST portalen och Kaldi-verktyget. Utförlig undersökning har gjorts av G.Salvi och N.Vanhainen i deras artikel "Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish"[4]. De använde sig av HTK (Hidden Markov Model Toolkit), MFCC särdragsextrahering och NST datasetet för att träna modellerna. Med detta recept lyckades de uppnå en 24% WER som de ansåg vara jämförbara med resultaten för ASR-system under liknande omständigheter [4].

Utöver G. Salvis och N. Vanhainens insatser har Emelie Kullmann [3] och Zimon Mossberg [5] utvecklat svenska ASR-system i samband med deras magistersavhandlingar. Kullmann använde även sig av NST dataseten men gjorde även jämförelser mellan GMM-HMM och DNN-HMM modeller, där DNN-HMM varianten producerade ett träffsäkrare system [3]. Mossberg konstaterar att utvecklingen av ett ASR-system med NST-datasetet och Kaldi, utan erfarenhet på expertnivå, leder till ett system med grova resultat [5].

### 6.1.2 Finsk undersökning

I och med att finska är det minsta av de stora nordiska språken (norska, svenska, danska, finska) är det inte överraskande att utbudet på tidigare undersökning är liten. Den främsta undersökningen i förbättring av finsk ASR utfördes, bland andra, av M. Kurimo och S. Enarvi et. al. i forskningarna "Modeling under-resourced languages for speech recognition" [28] och "Automatic Speech Recognition With Very Large Conversational Finnish and Estonian Vocabularies,"[29], båda utförda 2017.

I den senare forskningen lyckades de uppnå en 27,1% WER med hjälp av en blandning av NNLM (Neural Network Language Model) och delords n-gram modeller [29]. Detta var en avsevärd förbättring från deras tidigare resultat i 2016, då de endast uppnådde 48,4% WER.

Artikeln nämner även att det fanns rum för förbättring i användningen av NNLM, vilket är goda nyheter för framtida undersökning.

## 6.2 Förhinder

Som det har konstaterats i tidigare undersökningar [3][5][29] är en av de grundläggande hindren med utveckling av ASR-system för nordiska språk tillgängligheten av hög kvalitets röstdata. Avsaknaden av pålitlig röstdata tvingar forskare och utvecklare att utföra den krävande uppgiften att samla och testa data. Detta försvårar utvecklingsprocessen och höjer tröskeln för att påbörja utveckling av pålitlig ASR.

Ett problem karaktäristiskt för finsk-ugriska språk, bland annat finska, är att ett enstaka ord kan ha ett högt antal böjningar [28], exempelvis böjningen "taloineni" som motsvarar "med mitt hus" på svenska. På grund av denna karaktäristik är det svårt att hitta lämplig data som innehåller alla dessa sällsynta böjningar, och avsaknaden av dem orsakar stora problem för språkmodellen eftersom den inte längre kan pålitligt estimera böjningen av ett ord.

## 6.3 Förbättringskrav

Nordisk ASR är, åtminstone i forskningssyfte, i behov av stora förbättringar för att uppnå liknande resultat och engagemang som den engelskspråkiga ASR gemenskapen. Som det har tidigare nämnts i denna avhandlingen är avsaknandet av pålitlig data ett av de större problemen, vilket skulle kräva att organisationer, statliga och privata, skulle publicera delar av sin tillgängliga data för allmän användning.

Ett annat förbättringskrav är att mängden individer engagerade i forskning och utveckling av nordisk ASR skulle öka. Ett större nätverk av ASR forskare skulle fordra uppbyggningen av dataset och samtidigt undersöka vilka metoder vore mest passande för språket i fråga.

# 7. Diskussion

Källor:

[1]Haridas, Arul & Marimuthu, Ramalatha & Sivakumar, Vaazi. (2018). A critical review and analysis on techniques of speech recognition: The road ahead. International Journal of Knowledge-based and Intelligent Engineering Systems. 22. 39-57. 10.3233/KES-180374.

[2]Deng, li & Li, Jinyu & Huang, Jui-Ting & Yao, Kaisheng & Yu, Dong & Seide, Frank & Seltzer, Michael & Zweig, Geoff & He, Xiaodong & Williams, Jason & Gong, Yifan & Acero, Alex. (2013). Recent advances in deep learning for speech research at Microsoft. Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on. 8604-8608. 10.1109/ICASSP.2013.6639345.

[3] E. Kullmann Speech to Text for Swedish using KALDI

[4] Vanhainen, N., & Salvi, G. (2014). Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14). Retrieved from <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-158151>

[5] Z. Mossberg Achieving Automatic Speech Recognition for Swedish using the Kaldi toolkit

[6] Juang, B. & Rabiner, Lawrence. (2005). Automatic Speech Recognition - A Brief History of the Technology Development.

[7] P. Mermelstein (1976), "[Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*](https://books.google.com/books?id=wW9QAAAAMAAJ&q="Distance+measures+for+speech+recognition,+psychological+and+instrumental"&dq="Distance+measures+for+speech+recognition,+psychological+and+instrumental"&lr=&as_brr=0&as_pt=ALLTYPES&ei=zdRmSZjKLoH4lQTfqaXhBg&pgis=1)*,* C. H. Chen, Ed., pp. 374–388. Academic, New York.

[8] A. Mohamed Deep Neural Network acoustic models for ASR, University of Toronto, 2014

[9] Li Deng & Dong Yu, Deep Learning Methods and Applications p. 267-269

[10] Li Deng & Dong Yu, Deep Learning Methods and Applications p. 292-298

[11] Mozilla Common Voice Datasets <https://voice.mozilla.org/en/datasets>

[12] Google Speech-to-Text supported languages <https://cloud.google.com/speech-to-text/docs/languages>

[13] Dragon Transcription <https://www.nuance.com/dragon.html>

[14] Speechmatics Transcription <https://www.speechmatics.com/>

[15] Language models for Swedish authorities, AI Innovation of Sweden <https://www.ai.se/en/project/language-models-swedish-authorities> (hämtad 28.03.2020)

[16] Språkmodeller för svenska myndigheter Vinnova [https://www.vinnova.se/p/sprakmodeller-for-svenska-myndigheter/?\_t\_id=ZTx75B0kfSHqXVMeSGEJZA%3d%3d&\_t\_uuid=lx9lxF1JQ3usmphIP\_r\_IA&\_t\_q=RISE&\_t\_tags=language%3asv%2csiteid%3a6a0eda26-a5be-4f47-a778-b9393a63f812%2candquerymatch&\_t\_hit.id=Vinnova\_Models\_Pages\_ProjectPage/\_5e7123bb-b7d2-4a93-a26e-e2e9566d781b\_sv&\_t\_hit.pos=16](https://www.vinnova.se/p/sprakmodeller-for-svenska-myndigheter/?_t_id=ZTx75B0kfSHqXVMeSGEJZA%3D%3D&_t_uuid=lx9lxF1JQ3usmphIP_r_IA&_t_q=RISE&_t_tags=language%3Asv%2Csiteid%3A6a0eda26-a5be-4f47-a778-b9393a63f812%2Candquerymatch&_t_hit.id=Vinnova_Models_Pages_ProjectPage/_5e7123bb-b7d2-4a93-a26e-e2e9566d781b_sv&_t_hit.pos=16) (hämtad 28.03.2020)

[17] FAIA Finland’s Artificial Intelligence Accelerator <https://faia.fi/> (hämtad 28.03.2020)

[18] RISE Research Institutes of Sweden <https://www.ri.se/en>

[19] Peltarion <https://peltarion.com/>

[20] Kaldi ASR tookit homepage <http://kaldi-asr.org/>

[21] Kaldi ASR toolkit github <https://github.com/kaldi-asr/kaldi>

[22] Kaldi swedish recipe Emelie Kullmann <https://github.com/kaldi-asr/kaldi/tree/master/egs/sprakbanken_swe>

[23] Espresso ASR toolkit <https://github.com/freewym/espresso>

[24] Espresso: A fast end-to-end neural speech recognition toolkit, Y. Wang et.al. <https://www.researchgate.net/publication/335926466_Espresso_A_Fast_End-to-end_Neural_Speech_Recognition_Toolkit>

[25] LibriSpeech Resource page <http://www.openslr.org/12/>

[26] V. Panayotov, D. Povey et. al. Librispeech: An ASR corpus based on public domain audio books, <https://ieeexplore.ieee.org/document/7178964?section=abstract>

DOI**:** [10.1109/ICASSP.2015.7178964](https://doi.org/10.1109/ICASSP.2015.7178964)

[27] NST Språkbankens resurskatalog [https://www.nb.no/sprakbanken/#ticketsfrom?lang=en&query=alle&tokens=&from=1&size=12&collection=sbr](https://www.nb.no/sprakbanken/%22%20%5Cl%20%22ticketsfrom?lang=en&query=alle&tokens=&from=1&size=12&collection=sbr)

[28] Kurimo, M., Enarvi, S., Tilk, O. *et al.* Modeling under-resourced languages for speech recognition. *Lang Resources & Evaluation* **51,** 961–987 (2017). https://doi.org/10.1007/s10579-016-9336-9

[29] S. Enarvi, P. Smit, S. Virpioja and M. Kurimo, "Automatic Speech Recognition With Very Large Conversational Finnish and Estonian Vocabularies," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2085-2097, Nov. 2017.

Figures:

[Figure 1] Basic block diagram of a speech recognition system

E. Kullmann Speech to Text for Swedish using KALDI

URN: [urn:nbn:se:kth:diva-189890](http://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Akth%3Adiva-189890)

OAI: oai:DiVA.org:kth-189890

DiVA, id: [diva2:949757](http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A949757)