Automatisk Identifiering av Inandningspauser i Spontant Tal

- ett HMM/ANN-hybridsystem i Matlab

Lena Bystedt

lena.bystedt@cling.gu.se

December 2006

GÖTEBORGS
UNIVERSITET

Magisteruppsats i Datalingvistik, 20 poäng
Institutionen för Lingvistik, Göteborgs Universitet
Formell handledare: Torbjörn Lager

LUNDS
UNIVERSITET

Språk och litteraturcentrum,
Lingvistik/Språkteknologi, Lunds Universitet
Huvudhandledare: Johan Frid
Bihandledare: Anders Sjöström

Abstract

Automatic identification of inhalation pauses in spontaneous speech – hybrid HMM/ANN approach in Matlab

This thesis presents a system which has been implemented to satisfy a need in the research on how speech planning interacts with syntactic and prosodic structure in spontaneous speech. The long-term purpose of the research is to provide models for automatic parsing of spontaneous speech and for psycholinguistical modelling of speech production. Identification of inhalation pauses is an important step in the development of automatic methods for spontaneous speech parsing.

Identification of inhalation pauses is considered to be a keyword-spotting speech recognition problem. Hybrid HMM(Hidden Markov Models)/ANN(Artificial Neural Networks) approach is applied to this problem. Method gets 90,8% in Recall, 66,4% in Precision and 76,7% in F-score. Use of a threshold value for duration increases the F-score to 82,5%, therefore duration is considered to be relevant in performance optimization. Other proposed optimization parameters are better acoustic modelling, identification of the units causing false identifications prior to inhalation pauses identification and production of a more appropriate spontaneous speech corpus.

Sammanfattning

Automatisk identifiering av inandningspauser i spontant tal – ett HMM/ANN-hybrid-system i Matlab

Denna uppsats beskriver ett system som har implementerats för att tillgodose ett behov som har tillkommit under studier av hur planering av spontant tal samverkar med grammatisk och prosodisk struktur. Dessa studier syftar till att ge resultat som är relevanta för utvecklingen av modeller för automatisk parsing av spontant tal samt för den psykolingvistiska modelleringen av talproduktion. Identifiering av inandnings-pauser i spontant kontinuerligt tal är ett viktigt steg på vägen mot automatisk parsing av spontant tal.

Identifiering av inandningspauser betraktas vara ett nyckelords-identifieringsproblem och HMM(dolda Markov-modeller)/ANN(Artificiella Neurala Nätverk)-hybridsystem används istället för en tidigare metod som implementerar mönstermatchning. HMM/ANN-hybridsystem uppnår 90,8% i Recall och 66,4% i Precision, med sammanlagd bedömning på 76,7% i F-score. Applicering av ett tröskelvärde för duration ökar resultat i F-score till 82,5%, därför anses förbättrad modellering av duration vara användbar vid framtida systemoptimering. Övriga optimeringsförslag är bättre akustisk analys, identifiering och bortfiltrering av enheter som ofta identifieras felaktigt som ett försteg i systemet och effektiv framställning av ändamålsenlig korpus.

Förord

Det här arbetet är ännu ett exempel på en av de viktiga uppgifterna som datalingvisten och språkteknologen har - att leverera en lösning på ett språkligt problem som samtidigt möjliggör vidare forskning kring språkets natur.

Arbetet är utfört vid Lunds Universitet med finansiering från Göteborgs Universitet. Handledarna i Lund har varit Johan Frid och Anders Sjöström. Handledaren i Göteborg har varit Torbjörn Lager.

Johan Frid är forskarassistent vid Lunds Universitet, institutionen för Lingvistik/Språkteknologi, och disputerade 2003 med avhandlingen ”Lexical and Acoustic Modelling of Swedish Prosody”.

Anders Sjöström är forskningsingenjör vid Lunds Universitet, institutionen för Lingvistik/Språkteknologi, tidigare verksam i samma tjänst vid Lunds Universitet, institutionen för Matematik, avdelningen för numerisk analys.

Torbjörn Lager är professor i Allmän språkvetenskap och datalingvistik vid Göteborgs Universitet, och disputerade 1995 med avhandlingen ”A Logical Approach to Computational Corpus Lingustics”.

Anders Sjöström och Johan Frid bidrog mycket med litteraturtips och praktiska råd och lösningar under utvecklingens gång. De har haft en mycket förstående attityd och varit ett bra moraliskt stöd. Dessutom vill jag tacka Johan Frid för en mycket effektiv uppsatsgenomgång. Andra som har bidragit med sin kunskap i detta arbete är Merle Horne, professor i Allmän språkvetenskap, Lunds Universitet, Joost van de Weijer, forskarassistent, Lunds Universitet. Därtill tog Gilbert Ambrazaitis initiativ till att införskaffa talkorpusen som används i uppsatsen till Språk och Litteraturcentrumets (SOL) bibliotek vilket författaren är mycket tacksam för.

Jag vill också tacka min familj, och framförallt Fredrik, för gränslöst tålamod, stöd och till viss del påtryckningar under tiden för det här arbetet. Jag vill rikta ett tack till Dana Dannélls för synpunkterna på uppsatsen och positiv energi. Jag vill också tacka min opponent, Li Li, som har granskat det här arbetet och kommit med konstruktiva idéer och förslag. Dessutom vill jag tacka Zeki Hassan och Åsa Abelin för att ha uppmuntrat mina fördjupningar i fonetik.

Slutligen vill jag verkligen tacka Lunds Universitet, och framförallt SOL för att ha försett mig med en väldigt bra arbetsplats med utomordentliga verktyg, därbland databasen ELIN@Lund och SOL-biblioteket.

Innehållsförteckning

Illustrationsförteckning

Tabellförteckning

1 Introduktion

Detta avsnitt berättar om ämnet för denna uppsats och förklarar varför det kan vara intressant att automatiskt identifiera inandningspauser. Syfte och avgränsningar med arbetet presenteras. De viktigaste källorna som har använts för att uppnå syftet redovisas. Avsnittet avslutas med en beskrivning av uppsatsen upplägg.

Denna uppsats beskriver ett system som har tillkommit för att tillgodose ett behov som har uppstått under studier av hur planering av spontant tal samverkar med grammatisk och prosodisk struktur. Dessa studier syftar till att ge resultat som ”är relevanta för utvecklingen av modeller för automatisk parsing av spontant tal samt för den psykolingvistiska modelleringen av talproduktion”[1].

En av studierna (Horne et al., 2005a) tar upp en rad manuellt utförda observationer som utmynnar i en hypotes om tidsbegränsningar på talproduktionsenheter, också kallade för ”speech chunks” (Horne et al., 2005a) i spontant tal. Statistiska redogörelser saknas dock i denna studie. Författarna hämtar sina argument för tidsbegränsningen från minnesforskning och neurolingvistik. Materialet för observationerna är fältinspelningar från SweDia-projektet[2]; två kvinnliga talare från Götalanddelen[3].

Hypotesen består av följande antaganden

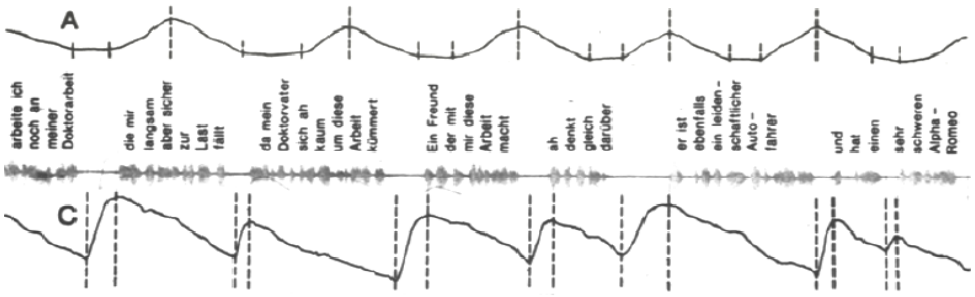
- en 2-2.5 sekunder talproduktionsenhet kan innehålla interna pauser
- en 2-2.5 sekunder talproduktionsenhet kan *inte* innehålla interna inandningspauser. Inandningspauserna förekommer bara på gränserna som omger talproduktionsenheterna
- en 2-2.5 sekunder talproduktionsenhet svarar optimalt mot en sats eller en fras.

1 http://www.ling.lu.se/projects/ProSeg2.html

2 http://swedia.ling.umu.se/info/info.html

3 Mailkorrespondens, Merle Horne, 20 oktober 2006

Eftersom talproduktionsenheterna antas förekomma mellan inandningspauser, kan identifiering av dessa vara ett första steg i att identifiera själva talproduktionsenheter. Talproduktionsenheterna antas svara mot grammatiska konstituenter. Därför kan uppstyckningen av spontant tal i dessa enheter vara användbart som ett försteg vid parsing av tal. Det har också uppmärksammats (Hird och Kirsner, 2002; McFarland, 2001; Winkworth, 1995; Ward, 1989; Conrad och Schönle, 1979; Henderson et al., 1965) att inandningspauser ofta förekommer vid grammatiska gränser i spontant tal, men att det varierar från 22% till 92% för de sex talarna i Winkworths undersökning (Winkworth, 1995). Conrad och Schönle visar i sitt arbete var inandningspausen förekommer i olika typer av tal. Figur 1 illustrerar hur en manlig talares bröstkorg expanderar och drar ihop sig under viloandning respektive andning under spontant tal. Skillnaden mellan dessa två är stor och vi ser att inandningspauserna förekommer vid strukturella gränser. Vi ser också att inandningspausen är ofta förekommande i tal.



Figur 1: Variationer i bröstkorgens omkrets (Y-axel) under tiden för (X-axel) viloandning (A) och andning under spontant tal (C), för en manlig talare. Inandningspauser är markerade med streckade linjer (Källa: Conrad och Schönle, 1979, sida 255, Figur 1 A-D).

Förutom detta specifika sammanhang kan modellering och identifiering av inandningspauser användas på andra sätt inom talteknologiska tillämpningsområden. Inom talsyntes skulle syntesens naturlighet kunna förbättras genom insättning av inandningspauser på ställen där en människa andas när denne talar (Sundaram och Naraynan, 2003). Även taligenkänningsprocessen skulle kunna förbättras genom tillägg av inandningspauser i språkmodellen så att systemet kan bortse från dem vid uppskattning av vad som har sagts (Englund, 2004; Weilhammer och Schiel, 1999; Schultz och Rogina, 1995; Ward, 1989).

Det bör nämnas redan nu att en metod för identifiering av inandningspauser utvecklades av Frid och Sjöström (Horne et al., 2005b) efter att den ovannämnda hypotesen hade tagit sin form. En grundlig utvärdering av metoden saknas i detta arbete, men författarna konstaterar att alla inandningspauser inte identifieras samt att många enheter som inte var inandningspauser blev identifierade som sådana. Därför åtog jag mig uppgiften att, på uppdrag av mina handledare och i konsultation med Merle Horne, utveckla en alternativ metod. En av avgränsningarna i uppsatsen är emellertid att metoderna inte ställs emot varandra vad gäller prestanda (se avsnitt 1.2).

Den här uppsatsen riktar sig till alla med intresse för talteknologi och grundläggande kunskaper i fonetik, maskininlärning och statistisk.

1.1 Syfte

Syftet med uppsatsen är att beskriva en automatisk metod för identifiering av inandningspauser i förinspelat spontant tal, samt att implementera, utvärdera och analysera densamma.

1.2 Avgränsningar

Metoden för identifiering av inandningspauser utvecklad av Frid och Sjöström (Horne et al., 2005b) som nämndes tidigare i kapitlet, ställs inte mot den metod som utvecklas här. Främsta orsaken till det är att den jämförelsen inte ryms inom uppsatsens ramar.

De antaganden inom hypotesen som ställdes upp tidigare i detta kapitel är baserade på ett talmaterial som är annorlunda än, men inte helt olikt, det talmaterial som används i uppsatsen. Därför är det möjligt att hypotesens antaganden inte är applicerbara på talmaterialet i uppsatsen (se avsnitt 4.1.1). Denna aspekt påverkar inte metodens utformning men gör att hypotesen inte kan prövas fullt ut. Ännu ett skäl är att fler behandlingssteg, utöver identifiering av inandningspauser, måste tas för att kunna identifiera talproduktionsenheter (Horne et al., 2005a).

Inandningspausernas akustiska och fysiologiska egenskaper presenteras endast kortfattat, och bara i syfte att möjliggöra för författaren att diskutera

kring övriga enheter i spontant tal som kan utgöra svårigheter för den automatiska identifieringen samt ge förslag på metodens vidareutveckling. Denna begränsning motiveras av att djupare kunskap om inandningspausens akustiska egenskaper kräver mer efterforskning. Därför väljs också en allmänt accepterad automatisk akustisk analys (se Tabell 3).

Den akustiska analysen resulterar i reducerad information inom vilken är endast de första enheterna möjliga att härleda till en viss akustisk egenskap. ANN avslöjar inte heller på ett uppenbart sätt vilka akustiska egenskaper som det har tagit fast på för att skilja en inandningpaus från icke-inandningspaus. Därför kan inte författaren på ett tillförlitligt sätt redogöra för vilka egenskaper hos inandningspausen som HMM/ANN-hybridsystemet har tagit fasta på för att skilja den från övriga enheter i tal.

De enheter som metoden felaktigt identifierar som inandningspauser analyseras inte eftersom det är ett tidskrävande manuellt arbete. Däremot föreslås en akustisk egenskap som verkar vara gemensam för dessa enheter och är lätt att få fram automatiskt. Exempel ges på hur denna kan användas som en säkerhetsfaktor (confidence factor[4]; Gold och Morgan, 2000, sida 59) för att få ner antalet felaktigt identifierade inandningspauser (se avsnitt 5.2.3 och 6.3).

Det färdiga systemet utgörs av en mängd funktioner där fokus har legat på att få fram en lösning för hur man kopplar samman HMM och ANN i utvecklingsmiljön Matlab. Systemet är därför inte riktat mot någon speciell användare eller någon särskild användning. Om man vill använda metoden i eget syfte får man alltså själv skapa överordnade funktioner.

1.3 Litteratur

Störst inflytande på metodens kärna, HMM/ANN-hybridsystemet, har Hervé Bourlards och Nelson Morgans publikationer haft. De har tillsammans med andra medarbetare arbetat med dolda Markov-modeller och lagt fram bevis (Renals et al., 1994) för hur de kan kombineras med artificiella neurala nätverk. Denna forskning är alltså grundläggande för metoden som används för att

4 Confidence factor eller confidence measure används inom nyckelordsidentifiering som ett ytterligare sätt att skilja de felaktigt identifierade orden från de korrekta. Detta används för att förbättra ett systems prestanda (Gold och Morgan, 2000).

identifiera inandningspauser.

Ramverket om dolda Markov-modeller och fördjupning i taligenkänning och närliggande områden har hämtats från Ben Golds och Nelson Morgans bok (Gold och Morgan, 2000) och John Holmes och Wendy Holmes bok (Holmes och Holmes, 2001). Information om maskininlärningsmetoder, inklusive artificiella neurala nätverk och utvärderingsmått kommer huvudsakligen från Ian Wittens och Eibe Franks bok ”Data Mining” (Witten och Frank, 2005). Svensk terminologi för ANN kommer från ett opublicerat kompendium vid Göteborgs Universitet, Filosofiska institutionen, med namnet ”Ett minne blott” av Helge Malmgren, professor i teoretisk filosofi.

Dokumentation till utvecklingsmiljön Matlab och dess inbyggda funktioner har haft ett starkt inflytande på metodens utformning och varit en källa till kunskap.

1.4 Disposition

Uppsatsens huvudsakliga del börjar med en introduktion i de kriterier enligt vilka svårigheten på en uppgift inom taligenkänning kan bedömas. En sådan bedömning ges för uppgiften att identifiera inandningspauser i spontant kontinuerligt tal. En fördjupning i svårigheterna med uppgiften görs via en beskrivning av inandningspausen samt genomgång av relaterat arbete.

Därefter följer kapitel tre som redogör för det teoretiska ramverket för metoden som används för identifiering av inandningspauser. Utvärderingsmetoder gås igenom översiktligt i samma kapitel. Det teoretiska ramverket är tänkt att fungera som stöd i efterföljande kapitel fyra som går igenom det tekniska genomförandet av metoden.

Kapitel fem rapporterar om det färdiga systemet samt hur det klarar av att identifiera inandningspauser. Diskussion kring systemets identifieringsförmåga, använd korpus och akustisk analys samt framtida arbete med hypotes kommer i kapitel sex. Uppsatsen avslutas med en koncis slutsats i kapitel sju.

2 Problembeskrivning

Detta avsnitt introducerar kriterier för bedömning av svårigheten för en uppgift inom taligenkänning. Givet dessa kriterier görs en bedömning av svårigheten för identifiering av inandningspauser. En fördjupning i problematiken görs i samband med att inandningspausen beskrivs utifrån akustiska och fysiologiska egenskaper. Dessutom belyses tidigare ansatser till automatisk identifiering av inandningspauser.

2.1 Uppgiftens svårighetsgrad

Ett taligenkänningssystems förmåga att klara av en viss uppgift kan presenteras i relation till svårighetsgraden på uppgiften det är ämnat att lösa. Gold och Morgan (Gold, Morgan, 2000, kapitel 5.3.1) presenterar en rad kriterier enligt vilka svårighetsgraden på en uppgift inom taligenkänning kan bedömas. Dessa är parafraserade nedan, ibland med författarens egna tillägg, där det lättare scenariot per punkt står först.

Talarberoende respektive talaroberoende

Klarar systemet av att känna igen tal från en eller flera individer. Fler individer introducerar större variabilitet i hur ett yttrande kan realiseras akustiskt vilket försvårar igenkänningsuppgiften. Även om många system kan hantera detta brukar problem uppstå med icke-modersmålstalare.

Isolerat respektive kontinuerligt tal

Känner systemet igen isolerade ord som följs av en tyst paus som tydligt markerar ordgräns, eller känner det igen kontinuerligt tal där orden inte avgränsas explicit med en sådan. I fallet med kontinuerligt tal kan systemet antingen känna igen godtyckliga satser i följd eller bara vissa på förhand utvalda ord bland dessa satser. Det senare fallet kallas från och med nu nyckelordsidentifiering.

Litet respektive stort lexikon

Ett större lexikon innebär generellt större utmaning eftersom större variabilitet introduceras med varje förekomst av samma fonem i olika lexikala sammanhang samt att fler ord kan förväxlas med varandra. Kriteriet kan emellertid vara missvisande eftersom ett litet lexikon kan innehålla enheter som är lika varandra och därför svåra att särskilja emellan.

Begränsad respektive mindre begränsad grammatik

Detta kriterium styrs av det ändamål som taligenkänningssystemet är skapat för. Ska systemet känna igen satser som produceras i kontakt med ett flygbiljettsbokningsystem, eller ska systemet känna igen vad som sägs i ett vardagligt samtal. Den senare är svårare på grund av att en sats kan formuleras på många fler sätt.

Läst respektive spontant tal

Läst tal är oftast mer hanterbart förutsatt att personen läser nerskrivet och på så sätt förplanerat material. Talet är då av jämnare talhastighet (stavelser/sekund) och innehåller färre eller inga typiska tecken för pågående planering av det som ska sägas härnäst. Spontant tal i en dialog, mer eller mindre planerad, är betydligt svårare att hantera. Det har större variation i talhastighet och fler disfluenser. Disfluenser är tvekljud som 'äää..am', att man stakar upp sig eller rättar sig när man talar.

Bra respektive dålig ljudkvalité

Miljö inom vilken talaren befinner sig samt mediet för överföring av tal kan påverka den akustiska analysen på ett positivt eller negativt sätt. Mindre frekvensomfång fångar inte detaljerna om vissa fonem vilket försvårar igenkänning av dem. Icke-språkliga ljud som sker parallellt med ett yttrande kan introducera för systemet okänd variabilitet.

Baslinje

Dessutom kan det färdiga systemets prestanda jämföras med en betydligt

enklare metod som man applicerar på samma problem. Är det lätt att lösa problemet med en enkel metod, så finns det inte i princip någon mening i att använda den svårare metoden. En baslinje talar också om hur lätt eller svår en uppgift är. Mer om vilken baslinje som används här finns i avsnitt 4.3.4.

2.1.1 Bedömning

Att identifiera inandningspauser som en enhet bland övrigt tal kan betraktas vara ett specialfall av nyckelordsidentifiering (engelska keyword-spotting). Särfallet med inandningspausen är att den varken är ett eller flera ord och inte heller ett fullvärdigt fonem. Antal enheter som ska kännas igen är alltså ett (inandningspausen), men i praktiken är detta antal två, inandningspausen och icke-inandningspausen. Den senare kallas för utfyllnad inom nyckelords-identifiering. Att endast en enheter ska kännas igen betyder att lexikonet för uppgiften är litet.

Talet som metoden är tänkt att fungera för, på grund av hypotesens utformning, är spontant. Därför är också data som den datadrivna metoden arbetar på av spontan natur. Talet som ska kännas igen är också kontinuerligt. Metoden tränas och utvärderas på olika talare, och är på så sätt talaroberoende, även om vissa talare introducerar fler problem än andra. Den färdiga metoden är dock begränsad till att fungera bäst på data liknande träningsdata. Inspelningen av data är gjord i studio och har bra ljudkvalité.

Att identifiera inandningspauser ger inte så komplicerad grammatik om man bara betraktar dem i form av bitar av ljud. Denna ”grammatik” modelleras av HMM/ANN-hybridsystemet. Den modellerar dock inte hur inandningspausen förhåller sig till icke-paus, om den senare skulle analyseras grammatiskt. Vi får alltså inte reda på, explicit, att i slutet på en sats kommer det en inandning. Detta ger enkel och begränsad grammatik.

Även om det bara handlar om att identifiera en enda enhet blir problemet relativt svårt eftersom inandningspausen kan förväxlas med andra (se avsnitt 2.2) liknande enheter i utfyllnaden. Flera författare konstaterar att det är problematiskt att identifiera enstaka enheter i tal, så som enskilda fonem (se avsnitt 2.3). Talet är spontant och kontinuerligt producerad av flera talare vilket försvårar uppgiften ytterligare. Den begränsade grammatiken gör uppgiften enklare och kvalitén på data förbättrar den akustiska analysen.

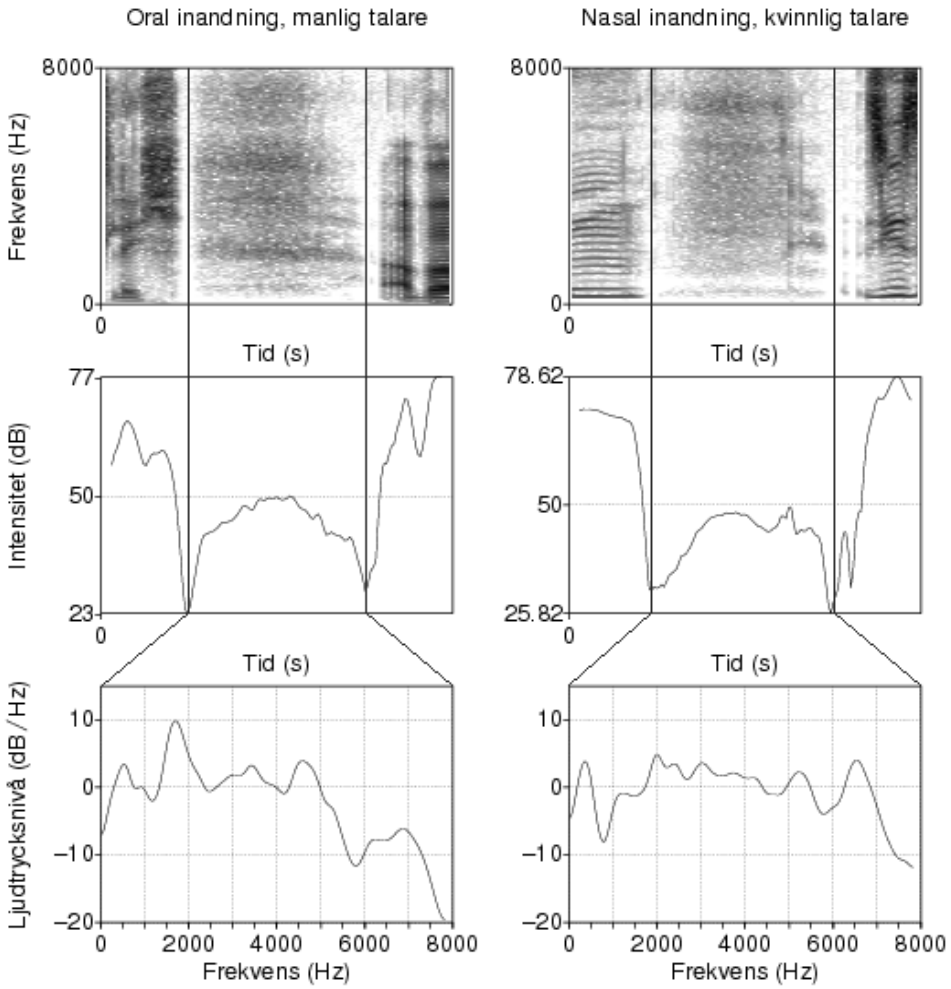
2.2 Inandningspausen

Den akustiska informationen i tal som identifieringsmetoden kan arbeta med representerar talets olika egenskaper. Till dessa hör bland annat duration, intensitet, energi vid de olika frekvenserna, utveckling av dessa i tiden samt samma information från omgivande kontext. Följande stycken beskriver trenderna i de ovan nämnda egenskaperna hos inandningspauser. För att relatera till svårigheten med uppgiften att identifiera inandningspausen redogörs även för de enheter i tal som kan innebära svårigheter för den automatiska metoden.

Andningscykeln består av en inandnings- och utandningsfas (Hixon, 1973), vilket också framgår av Figur 1 (A). Andningscykeln är oregelbunden i spontant tal (Hixon, 1973) i jämförelse mot viloandning Figur 1 (C). Märk väl att Figur 1 illustrerar en monolog. I en dialog kan man anta förekommer både oregelbunden och regelbunden (viloandning) andningscykel.

Följande beskrivning är baserad på Figur 2 nästa sida. Inandningspausen är tonlös och har en operiodisk ljudkälla. Den nasala inandningspausen följs ofta av att man öppnar munnen med ett klickljud. Intensiteten i inandningspausen, i relation till det mänskliga hörseltröskelvärdet och ungefär 0,5 meter från mikrofonen, verkar generellt sett vara lägre än det producerade talet. Enligt Wikipedia[5] ligger intensiteten för andning på 10 dB, mätt från 3 meter (jämför med 50-70 dB – normal talnivå). Enligt Conrad och Schönle varierar durationen på inandningspauserna från cirka 0,4 sekunder till 0,8 sekunder (av författaren tolkat värde från Figur 4; Conrad och Schönle, 1979) i spontant tal i skillnad mot att vara mer stabil under läst tal och viloandning (Figur 4; Conrad och Schönle, 1979, sida 258).

5 http://en.wikipedia.org/wiki/Decibel



Figur 2: Exempel på två typer av inandningspauser: oral, manlig talare; nasal, kvinnlig talare.

Givet denna beskrivning av inandningspauser kan det förmodas att enheter som ligger i riskzonen för förväxling med en inandningspaus är ljud som också har en operiodisk ljudkälla. Detta gäller tonlösa frikativor och tonlösa klusiler. Samtliga klusiler kan vara aspirerade vilket sker när luftströmmen fortfarande pågår innan den eventuellt tonande källan sätter igång. Tidsförloppet innan den tonande källan startar kallas Voice Onset Time (VOT). Författarens förmodan är emellertid att den del av VOT som metoden skulle kunna ta fast på inte pågår lika länge som en inandningspaus och därför kan duration användas för att

filtrera bort dessa enheter. Vi ser också (Figur 2, spektrogram) att energifördelningen vid de olika frekvenserna för frikativor är högre än för inandningspausen. Utandningar introducerar en särskild problematik på grund av korpusens utformning (se avsnitt 4.1.1 och 6.2).

Röstkvalitéerna som kan utgöra problem är läckande röst samt om man viskar istället för att tala högt. Författaren tror också att ingressivt tal[6] kan utgöra problem för identifieringen av inandningspauser.

2.3 Relaterat arbete

För att få en uppfattning om hur svårt problemet att identifiera inandningspauser är kan man också undersöka vad andra har gjort tidigare. För att få en någorlunda heltäckande bild kan identifiering av inandningspausen betraktas från olika synvinklar. De synvinklarna som utökas nedan är: den fysiologiska, den andningsinriktade – inandningspaus och utandning är samma sak, den inandningspausinriktade – där man riktar sig in på att identifiera just inandningspauser, den foneminriktade – en inandningspaus är ett fonem och den perceptionsinriktade – hur väl människor klarar av att identifiera inandningspauser.

Den fysiologiska synvinkeln ger oss accepterade verktyg för att identifiera inandningspauser, ett exempel är en pneumograf (se Figur 1, kapitel 1 och avsnitt 6.2). Dessa verktyg är direkt oanvändbara eftersom de kräver att den personen, vars tal man ska känna igen, är fysiskt närvarande. Fallet med automatisk taligenkänning är att endast talsignalen är tillgänglig. Vi lämnar dessa för tillfället men återkommer till dem i diskussionen kring korpustillverkning (se avsnitt 6.2).

Wightman och Ostendorf poängterar vikten av inandningspauser (Wightman och Ostendorf, 1994, sida 472) med övergår nästan direkt till att resonera kring andning och gör alltså ingen skillnad i något av arbeten på inandningspaus och utandning. I det tidigare arbetet (Wigthman och Ostendorf, 1991) applicerar författarna en HMM (mer om HMM i avsnitt 3.1) med normalfördelningar för att presentera akustisk information i modellen. De

6 Eklund, R., (2004). ”Disfluency in Swedish human-human and human-machine travel booking dialogues”, Linköpings universitet.

uppnår 72,3% (Wightman och Ostendorf, 1991, sida 322) i antal korrekt identifierade andningar. Talmaterialet består av tre nyhetsstycken upplästa av ett professionellt nyhetsankare (Wightman och Ostendorf, 1991). Detta sätt (arbetet från 1991) att identifiera andningar i uppläst tal sägs av författarna ha vidareutvecklats till en treklass-klassificerare med samma underliggande metod som tidigare och att ”(an extension of [38]) detected all 364 breaths in 63 paragraphs of speech, with a 3% false insertion rate.” (Wightman och Ostendorf, 1994, sida 472). Mer information om vilka tre klasser det rör sig om, och hur de har gått tillväga för att utöka HMM finns inte i artikeln. Det gick alltså enligt författarna mycket bra för treklass-HMM med normalfördelningar att detektera andningspauser i nyhetsuppläsningar. De problem som spontant kontinuerligt tal från flera olika talare introducerar hade förmodligen visat på svagheterna med denna metod. Därför kan vi samtidigt sluta oss an till att en metod som identifierar inandningspauser i spontant kontinuerligt tal kommer att prestera sämre än för läst tal.

Att modellera andning som en typ av övriga ljud representeras av två arbeten (Schultz och Rogina, 1995; Ward, 1989). Schultz och Rogina arbetar med talmaterial som är högst aktuellt här, nämligen en talkorpus som är baserad på tal från människa-till-människa-interaktion. Svårigheten med den taligenkänningsuppgiften är att författarna vill särskilja mellan ”icke-språkliga” mänskliga ljud inklusive andningar och övriga externa ljud. Den högsta korrektheten de uppnår är 66% för engelska och 70% för tyska (Schultz och Rogina, 1995, sida 295). Författarna konstaterar att många ljud förväxlas med varandra. Dessa resultat bekräftar resonemang om att en inandningspaus kan förväxlas med många andra enheter i spontant kontinuerligt tal (se avsnitt 2.2). Procentantal i Schultz och Roginas arbete kan *inte* jämföras mot HMM/ANN-hybridsystemet eftersom syfte och utvärderingsmetoderna skiljer sig. Ward (Ward, 1989) inkorporerar andning i sina modeller för taligenkänning, men tyvärr behandlas inte andningen som någon speciell enhet utan som övrigt ljud som modelleras bort.

Vad gäller att identifiera den självständiga inandningspausen i kontinuerligt spontant tal är det endast, vad författaren vet i dagens läge, Horne, Frid och Sjöström (Horne et al., 2005b) som har tagit sig an denna uppgift. Metoden i detta arbete är mönstermatchning där en mall av en inandningspaus jämförs stegvis mot ett längre målljud och tröskelvärden används för att avgöra om en inandningspaus har identifierats. Tyvärr saknas en fullständig utvärdering av

metoden i detta arbete men författarna konstaterar att högt antal korrekt identifierade inandningspauser fås på bekostnad av många felaktigt identifierade inandningspauser. Författarna lägger märke till att det generellt går bättre att identifiera inandningspauser i manligt än kvinnligt tal. Författarna utlyser ytterligare metoder för reduktion av antal felaktigt identifierade inandningspauser (Horne et al., 2005b). Förslag på en sådan metod ges i avsnitt 5.2.3 och 6.3 eftersom samma problem uppstår för HMM/ANN-hybridsystemet.

Ett annat arbete implementerar ANN för att identifiera inandningar och utandningar hos sovande personer. Två separata nätverk används för ändamålet. 98% av alla inandningspauser identifieras korrekt (Sá, 2002, abstract). Andningsmönstret i det talmaterialet borde vara likt viloandning i Figur 1. Svårighetsgraden på denna uppgift är nästintill minimal enligt kriterierna uppställda i avsnitt 2.1 och korrektheten är motsvarande. Arbetet är fysiologiskt inspirerat.

Inandningspausen skulle tentativt kunna likställas med fonem trots att de språkliga funktioner hos dessa båda enheter antagligen är helt skilda. Flera arbeten konstaterar att svårighetsgraden är hög vad gäller att identifiera enstaka fonem i kontinuerligt spontant tal från många talare. Närmast inandningspausen ligger fonem med operiodisk ljudkälla - frikativor och klusiler (se avsnitt 2.2). Abdelatty Ali och kollegorna (Abdelatty Ali et al., 1998) beskriver fyra kategorier av fonem som deras metod skall särskilja på: ”sonoranter, klusiler, frikativor och tystnader” (Abdelatty Ali et al., 1998, (abstract) sida 118). De uppnår höga resultat i korrekthet 86% för klusiler och 90% för frikativor. Den stora fördelen med deras arbete är att de skapar ett ramverk av en välvald och optimerad akustisk analys. Deras förslag är att ramverket skall användas i kombination med HMM och ANN (Abdelatty Ali et al., 1998). Författarna använder bland annat det akustiska spektrala särdraget ”center of gravity” som av ett annat arbete (Gordon et al., 2002) anses vara användbart vid särskiljning av frikativor. Abdellatty Alis arbete kan och bör användas i framtida utveckling av identifiering av inandningspauser (se avsnitt 6.2). Deras arbete används inte här.

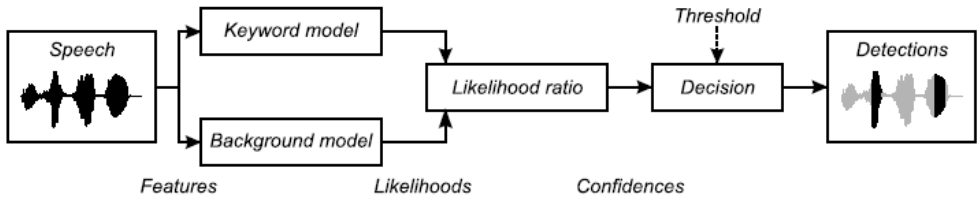
Den perceptuella sidan av identifiering av inandningspauser har undersökts av Iwarsson (Iwarsson, 2000). Denna studie består i att trettioen studenter i patalogi får lyssna till läst tal producerad av tolv talare. Totalt blir 91,3% av inandningspauserna identifierade av dessa lyssnare fast med stor variation. Författaren konstaterar att inandningspauser förekommer huvudsakligen innan början av en sats. De inandningspauserna som avviker från detta mönster är

också de som till stor del förbises av lyssnarna. Andra faktorer som påverkar den perceptuella identifieringen av inandningspauser negativt är lyssnarnas antagande om hur ofta en person vanligtvis andas in, hur luftvolymen förändras vid inandningspausen samt ljudnivån på inandningspausen (Iwarsson, 2000). Detta experiment visar på att det kan vara svårt även för tränade mänskliga lyssnare att identifiera inandningspauser i läst tal. Låg ljudnivå är en av orsakerna vilket är samstämmigt med tidigare beskrivning av inandningspausen (se avsnitt 2.2). Spontant kontinuerligt tal måste innebära en ännu större utmaning, därför tror författaren här att det kommer att vara svårt för HMM/ANN-hybridsystemet att uppnå 91% vad gäller antal träffade inandningspauser samtidigt som att antal felaktigt identifierade enheter är lågt.

Till sist vill författaren påpeka att inget av dessa arbeten tar upp möjligheten att ett aktivt andningsmönster kan åtföljas av viloandning då talaren förvandlas till en lyssnare i en dialog. Denna aspekt bör uppmärksammas i framtida arbeten.

3 Bakgrund

Nyckelordsidentifiering (Szöke et al., 2005) kan genomföras med hjälp av mönstermatchning eller dolda Markov-modeller som identifieringsmetod. Hur nyckelordsidentifieringssystem generellt ser ut illustreras av Figur 3.



Figur 3: Allmän systemöversikt för nyckelordsidentifiering (Källa: Szöke et al., 2005)

Arbetet som har gjorts här implementerar dolda Markov-modeller, där de traditionella sannolikhetsfunktioner som förknippas med varje tillstånd i modellen har ersatts med skalade betingade sannolikheter från ett artificiellt neuralt nätverk. Just därför kallas denna metod för HMM/ANN-hybridsystem.

Nedan beskrivs HMM, ANN och hur dessa kan kombineras. Utöver det beskrivs hur den akustiska analysen går till samt de mått som används för att utvärdera HMM/ANN-hybridsystemets identifiering av inandningspauser.

3.1 HMM

Dolda Markov-modeller har på senare tid i allt större utsträckning använts inom taligenkänning. Deras styrka ligger i att de kan känna igen en sekvens av exempelvis fonem, trots att de akustiska realisationerna för dessa varierar beroende på vilken kontext de förekommer i. Man representerar denna varians med hjälp av sannolikhetsfördelningar för de enheter man är intresserad av. Dessa fördelningar används sedan för att ge den mest troliga uppskattningen av fonem- eller ordsekvens som har yttrats givet en representation av tal.

Anledningen till att Markov-modellerna lämpar sig så väl för sekvensigenkänning är deras konstruktion. En Markov-modell är en finit tillståndsautomat där varje tillstånd, exempelvis beteckning för ett fonem, är

riktat mot andra tillstånd i modellen inklusive sig själv via övergångssannolikheter. Dessa bestämmer hur modellen kan ta sig genom tillstånden och därmed generera en sekvens av tillstånd. Övergångsannolikheterna beräknas i form av relativ frekvens för övergångarna mellan tillstånden. Om en viss övergångssannolikhet är större än de övriga, kommer modellen oftare att gå den vägen. Sannolikheten för en viss sekvens av tillstånd är produkten av övergångssannolikheter som man får när man passerar just dessa tillstånd.

För att göra Markov-modellen användbar för taligenkänning måste man förknippa akustisk information med varje tillstånd. När man gör det så ser man inte längre själva tillstånden, utan istället en sannolikhetsfunktion, vanligtvis flera normalfördelningskurvor, som talar om hur stor chansen är att viss akustisk information kommer att erhållas när modellen är i tillståndet, så kallade emissionssannolikheter.

Dessa sannolikheter, tillsammans med övergångssannolikheter, gör att vi kan uppskatta vilken sekvens av tillstånd som har ägt rum givet någon akustisk information. Oftast resulterar det i flera förslag på troliga sekvenser, därför används en metod, Viterbi-uppskattning[7], för att plocka fram den mest troliga sekvensen.

Problemet kan uttryckas på följande sätt: P(w|Y) som är posteori sannolikhet för ett ord (w) givet ”observerad” akustisk information (Y). P(w|Y) kan vara svårt att beräkna rakt av därför får man använda sig av Bayes sats som underlättar beräkningen. Då kan man likställa problemet med att beräkna P(w|Y) = P(Y|w)P(w)/P(Y). Man räknar då sannolikheten för en viss akustisk ”observation” givet ett ord (emissionssannolikheterna), gånger sannolikheten för ett ord genom sannolikheten för en akustisk observation. Sannolikheten för akustisk observation är konstant. Därför kan man bortse från denna sannolikhet i beräkningen eftersom den inte bidrar till resultatet. Vi reducerar formeln till P(w|Y) = P(Y|w)P(w). Just P(Y|w) kan räknas fram på olika sätt, flera normalfördelningar eller neurala nätverk.

7 http://en.wikipedia.org/wiki/Viterbi_algorithm

3.2 ANN

Articifiella neurala nätverk kan användas för att förknippa akustisk information med tillstånd i en Markov-modell. Artificiella neurala nätverk är en biologiskt inspirerad maskininlärningsmetod baserad på neural teori och är bland annat bra på att lära sig generalisera och att känna igen mönster i data. Artificiella neurala nätverk kan givet en bit akustisk information uppskatta, i form av en sannolikhet, till vilken klass (exempelvis beteckning för ett fonem) denna bit av information tillhör. Skillnaden mellan neurala nätverk och dolda Markov-modeller är att det neurala nätverkverket ger en uppskattning för endast en liten bit av tal, så kallade lokala sannolikheter. En dold Markov-modell med Viterbi-algoritmen ger däremot den mest troliga uppskattningen för en sekvens av sådana bitar av tal, på så sätt får man in tidsapekten i modellen. Nätverket ser bara ett steg av ljud i taget och kan gissa på en felaktig klass om denna råkar dela vissa akustiska särdrag med den korrekta klassen. Dolda Markov-modellen kompenserar just för denna effekt genom att titta på kontexten som biten befinner sig i och uppskatta huruvida det är en rimlig identifikation eller inte.

Neurala nätverk är ett sätt att representera ett funktionssamband mellan indata och utdata. I det här arbetet används neurala nätverk för mönsterklassificering. Det finns olika varianter på neurala nätverk samt inlärningsmetoder för olika typer av uppgifter.

En vanlig variant för mönsterklassificering är 'feed-forward'-nätverk. 'Feed-forward' innebär att flödet av informationen genom nätverket är enkelriktat. Komponenterna i ett 'feed-forward'-nätverk är neuroner och aktiveringsfunktioner i dessa samt anslutningar, med tillhörande vikter, mellan dessa neuroner. Neuroner kan vara uppställda i lager med en eller flera neuroner där varje neuron är ansluten till varje neuron i nästa lager. Ett 'feed-forward'-nätverks kraftfullhet kan bestämmas genom valet av antal lager av anslutningar samt antal neuroner i lager med neuroner.

Neurala nätverk måste först tränas för att därefter kunna användas för validering och test. Det som skiljer dessa faser åt, är att under träningen så anpassas vikterna i anslutningarna mellan neuronerna för att passa in på utdata. Men under valideringss- och testfasen är vikterna frysta. Dessutom är datamängderna separerade så att samma data inte förekommer i olika mängder.

Anledningen till att man väljer neurala nätverk istället för flera normalfördelningar för att uppskatta emissionssannolikheter är att neurala

nätverk erbjuder mycket bra urskillningsförmåga. Det är också lätt att ändra nätverkets struktur och experimentera med olika typer akustisk information. Oftast krävs det färre parametrar och mindre träningsexempel än för uppskattning av flera normalfördelningar. Dock är det fortfarande så att då det rör sig om uppgiften att känna igen kontinuerligt tal med obegränsat lexikon presterar traditionella dolda Markov-modeller bättre än HMM/ANN-hybridsystem.

3.3 HMM/ANN-hybridsystem

Avsnittet bygger på Renals och kollegornas arbete ”Connectionist Probability Estimators in HMM Speech Recognition” (Renals et al., 1994).

För att inkorporera neurala nätverk, som emissionssannolikheter P(Y|w), i Markov-modeller, och därmed skapa en HMM/ANN-hybridsystem krävs en omräkning av det svar som ANN matar ut givet en bit akustisk information. Svaret är en a posteori sannolikhet mellan 0 och 1 som är nätverkets klassificering av den akustiska informationen i en kategori. Ju närmare sannolikheten är 1 desto större chans att det exempelvis rör sig om en inandningspaus.

Vi har nu en remsa med a posteori sannolikheter för varje bit akustisk information. Som nämndes tidigare är dessa lokala sannolikheter inte tillräckligt bra för att beskriva en sekvens av tillstånd. Detta kan kringgås genom applicering av Bayes regel på varje sannolikhetsvärde i remsan. Vilket innebär att vi dividerar varje a posteori sannolikhet i remsan med a priori sannolikhet för varje tillstånd. Formeln för det ser ut på följande vis:
$P(Y|w) = P(w|Y)P(Y)/P(w)$. Det vi får ut, vänstra argumentet i föregående formel, är emissionssannolikheter som talar om hur stor chansen är att vi kommer att erhålla en viss symbol, eller akustisk information, när vi befinner oss i ett visst tillstånd. Dessa sannolikheter används i den dolda-Markov modellen där Viterbi-algoritmen uppskattar den bästa vägen genom tillstånd. För närmare information och exempel på hur HMM/ANN-hybridsystemet är implementerat och anpassat i syfte att identifiera inandningspausen se avsnitt 4.3.3.

3.4 Akustisk analys

För att automatiskt kunna karaktärisera ljud under en viss tid måste det presenteras digitalt utifrån relevanta akustiska egenskaper. Ett steg i processen är att stega genom ljudet och applicera en funktion på varje steg. Denna process kallas för fönstring. Bredden på fönstret bestämmer hur bra upplösning man får i frekvens- och tidsled. Ett kortare fönster ger bra upplösning i tid och en bra uppskattning av spektrumenvelopp (formantstruktur). Ett längre fönster ger en bättre upplösning i frekvens och är bland annat bra för att skilja ut deltoner[8].

Vid varje fönstringsmoment hämtas information om hur energin i ljudet fördelar sig vid de olika frekvenserna. Det resulterar i ett spektrum. Information i ett spektrum kan reduceras till ett cepstrum där betydligt färre komponenter, så kallade cepstrala koefficienter, kan användas för att uttrycka samma information som i ett spektrum (Elenius och Blomberg, 2000). Samtidigt som information från spektrumet reduceras till cepstrum separeras akustiska egenskaper, och kan därför för sig beskriva en viss akustisk egenskap. Nollte koefficienten representerar energi i ljudet, för det specifika steget man står i, opåverkad av omgivande energi. De högre koefficienterna står för mer detaljerad information om spektrumets utseende. Vanligt antal använda cepstrala koefficienter är 12. Det finns också möjlighet att uttrycka dessa koefficienternas utveckling i tiden med delta som representerar hastighet och delta-delta som representerar acceleration. På så sätt fångas talets dynamiska struktur.

Ofta vill man också skala om den akustiska informationen enligt frekvensuppfattningen i ett mänskligt öra. Anledningen till det är att språkliga ljud ligger just i det område där mänsklig hörsel är känsligast. I detta fall appliceras en mängd filter, Melfilter, som skalar om energierna och frekvenserna enligt Melskala innan man räknar fram ett cepstrum.

8 http://www.speech.kth.se/courses/1120/2006/frl2_2f1120.pdf

3.5 Utvärdering

Mönsterklassificering, det vill säga hur väl metoden har lyckats lära sig att ett visst mönster tillhör en viss kategori, kan bedömas i form av alternativ uppställda i Tabell 1. HMM/ANN-hybridsystemet har två kategorier att välja emellan, 0 för icke-inandningspaus och 1 för inandningspaus.

Dessa två alternativ leder till fyra identifieringsalternativ: 1. en inandningpaus har identifierats som en inandningspaus; en inandning har identifierats som en icke-inandning; 3. en icke-inandning har identifierats som en icke-inandning; 4. en icke-inandning har identifierats som en inandning.

Tabell 1: Olika utfall vid tvåklassprediktion.

		Predicerad klass	
		1	0
Faktisk klass	1	Sann Positiv	Falsk Negativ
	0	Falsk Positiv	Sann Negativ

1 – inandningspaus, 0 – icke-inandningspaus. Adapterad från Witten, 2005, sida 162.

Genom att kombinera alternativen Sann Positiv (SP), Falsk Negativ (FN), Sann Negativ (SN) och Falsk Positiv (FP) på olika sätt kan man få fram olika mått som talar om systemets förmåga att identifiera rätt kategori givet en bit av ett ljud. De kombinationerna som används här är SannPositiv-värde (engelska ”True Positive Rate”), FalskNegativ-värde (engelska ”False Negative Rate”), SannNegativ-värde (engelska ”True Negative Rate”), FalskPositiv-värde (engelska ”False Positive Rate”), TotalFramgång-värde (engelska ”Sucess Rate”). Dessa värden ligger mellan 0 och 1. Tabell 2 visar hur dessa mått räknas fram utifrån de möjliga prediktionsutfall i Tabell 1.

SannPositiv-värdet talar om hur många inandningspauser har identifierats som inandningspauser. Ju högre värde desto bättre.

FalskNegativ-värdet talar om hur många inandningspauser har identifierats som icke-inandningspaus. Ju lägre värde desto bättre.

SannNegativ-värdet talar om hur många steg i icke-inandningpaus har identifierats som icke-inandningspaus. Ju högre värde desto bättre.

FalskPositiv-värdet talar om hur många steg i icke-inandningspaus som systemet har identifierat som inandningspaus. Ju lägre värde desto bättre.

TotalFramgång-värdet talar om hur många steg totalt har identifierats rätt. Ju högre värde desto bättre.

När bedömningen av en metods identifieringsförmåga går från stegnivå till ordnivå lämpar sig andra mått bättre. De mått som kan användas i fallet med bedömning av nyckelordsidentifiering är mått som används vid informationsåtkomst (engelska ”information retrieval (IR)) (McOwan et al., 2005). Dessa presenteras längre ner.

Tabell 2: Utvärderingsmått för bedömning av stegklassificering.

Värde	Formel
SannPositiv 1 som 1	(SP/(SP+FN))*100%
FalskNegativ 1 som 0	100-SannPositiv
SannNegativ 0 som 0	(SN/(SN+FP))*100%
FalskPositiv 0 som 1	100-SannNegativ
TotalFramgång	((SP+SN)/(SP+FN+SN+FP))*100%

1 – inandningspaus, 0 – icke-inandningspaus. Adapterad från Witten, 2005, sida 162-163.

Recall – (SP/(SP+FN))*100%, hur många av inandningspauser i facit har metoden identifierat.

Precision – (SP/(SP+FP))*100%, hur många av de identifierade enheterna faktiskt var inandningspauser.

F-score – (2*Recall*Precision)/(Recall+Precision), ett värde som väger samman Recall och Precision. Används för att få en sammanlagd bedömning.

Utöver dessa mått kan man fördjupa sig i de identifierade enheternas egenskaper. Framförallt för att få en insikt om vad det är i identifieringsprocessen som gått snett. Den egenskapen som undersöks här är duration (se avsnitt 5.2.3).

4 Implementation av metod

4.1 Kiel Corpus of Spontaneous Speech

Material som används för att utveckla den datadrivna metoden för identifiering av inandningspauser är en korpus som har gjorts vid Kiels Universitet. Volymerna som används är 1 och 2.

Talet i korpusen är tyska med olika dialekter. Dialogen sker enligt ett scenario, sådant att två personer avtalar ett möte utifrån en kalender de har fått. Talarna måste trycka på en knapp för att få tala. Därför uppnås en strikt uppdelning av talarturen. Målet med detta scenario är att efterlikna människa-till-maskin interaktion. Information om talarnas kön, ålder och dialekttillhörighet finns i Appendix B.

Inspelningen sker i ett ljudisolerat rum. Samplingsfrekvensen är 16000 Hz, vilket gör att det analyserbara frekvensintervallet ligger mellan 0 och 8000 Hz. Upplösningen på ljudet är 16 bitar, vilket innebär att mycket information om den ursprungliga talsignalen bibehålls.

Exempel på hur en annoteringsfil till varje talares tur ser ut finns i Appendix A. Informationen som används för att skapa träningsdata till metoden beskriven i avsnitt 4.3.2 är den som följer markören 'hend' i en sådan fil. Informationen består av sampelnummer, vilket markerar position i ljudet, och motsvarande markör. Markörena som används för att markera att en inandningspaus eller en utandning är närvarande är följande: ’#h:’,’#-h:’,’$h:’, ’$-h:’. Markören som föregås av '$' innebär att enheter förekommer inom ett ord. För mer detaljer rekommenderas korpusens dokumentation (se Kohler et al., 1995).

Av stycket innan framgår att korpusen inte gör någon skillnad på en inandningspaus och en utandning. Det blev känt när valet av korpus redan var bestämt, författarens förhoppning var att en utandning skulle vara mindre frekvent än en utandning och att metoden skulle på något se bortse från dessa. Tyvärr visar det sig att metoden har lärt sig just inandningspausen och utandning. Mer om detta i diskussionskapitlet (se avsnitt 6.2). Användningen av enheter som förekommer inom ett ord kan tyckas vara ett tveksamt val, eftersom man framförallt vill identifiera inandningspauser som befinner sig vid

strukturella gränser. Valet orsakas av att författaren inte vill introducera inkonsistens i representation av mönster och facit till det artificiella neurala nätverket. Att talet är tyska betyder ingenting för metodens utformning, men det påverkar vilket språk metoden kan känna igen. Dessutom introducerar talad tyska vissa svårigheter eftersom många frikativa ljud finns och klusiler är mycket aspirerade (Jessen M., 1999).

4.2 Matlab

För att utveckla systemet har jag jobbat i utvecklingsmiljön Matlab[9]. Fördelen med miljön att det går snabbt att implementera ett koncept eftersom många funktioner är redan implementerade. Dessa funktioner är välarbetade med mycket kringinformation och effektiv dokumentation. Programmeringsspråket är imperativt, grundstrukturerna för data är verktorer och matriser.

Matlab kan utökas med olika verktygslådor som innehåller funktioner för att lösa specifika eller mer generella problem. De verktygslådorna som jag har använt mig av i implementationen är Statistic Toolbox, Neural Network Toolbox och Voicebox.

4.3 Tillvägagångssätt

4.3.1 Uppdelning av korpus

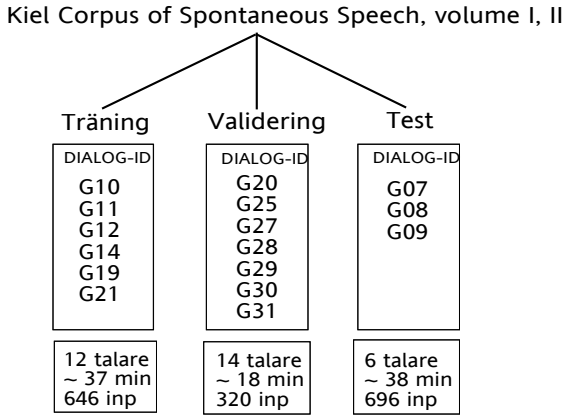
Materialet i en korpus måste delas upp i oberoende mängder för att ge en rättvis uppskattning av en metods förmåga att utföra en viss uppgift.

I Figur 4 på nästa sida finns en schematisk uppställning över hur talmaterialet från Kiel Corpus of Spontaneous Speech, volymerna ett och två, har delats upp i mängderna: träning, validering och test. I Appendix C finns en detaljerad lista över samtliga filer.

Fördelningen mellan de olika mängderna, räknat utifrån antal inandningspauser: 39% träningsdata, 19% valideringsdata och 42% testdata (se

9 http://www.mathworks.com/

Figur 5). Tränings- och valideringsdata utgör tillsammans 58%. Orsaken till att författaren har valt denna uppdelning ligger i: hur talmaterialet är beskaffat; hur den akustiska analysen är uppbyggd; hur träningsproceduren för ANN går till; minnesbegränsningar på datorhårdvaran.



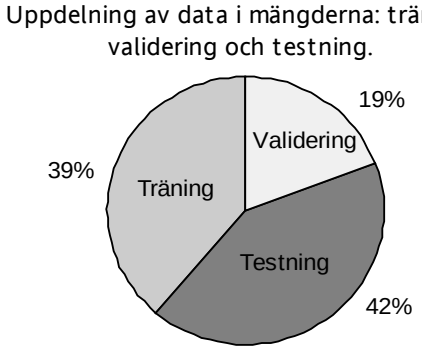
Figur 4: Uppdelning av filerna i Kiel Corpus of Spontaneous Speech, volymerna ett och två, i mängderna träning, validering och test, sorterade efter dialog ID. "inp" står för inandningspauser.

Eftersom författaren ville uppnå separation mellan datamängderna även på talarnivå, bestämdes uppdelningen av dialog ID. Dialog ID är unikt för varje par av talare i korpusen. Antal filer i de olika dialogerna som finns i Kiel Corpus of Spontaneous Speech volym I och II är olika därför är valideringsdatamängden mindre än de övriga datamängderna. Författaren ville ge systemet störst utmaning under valideringsfasen för att undvika optimering av systemet mot ett litet antal talare, vilket förklarar det relativt höga antalet talare i valideringsmängden.

Den akustiska analysen, som diskuteras i nästa avsnitt, generar cirka 320 Megabyte (Mb) data (datafilen sparad i ASCII-format[10]) för träningsdata och cirka 150 Mb för valideringsdata. Eftersom träningsalgoritmen förutsätter att data bearbetas satsvis och optimal generaliseringsförmåga kräver valideringsdata, körs båda datamängderna, cirka 470 Mb, efter varandra för

10 Ett sätt att spara data för att göra det tillgängligt på många andra datorsystem.

varje träningsiteration. Författaren uppskattar det genomsnittliga antalet iterationer, i medelvärde och för olika stora neurala nät per träningsession, till sextiofem. Denna träningsprocess är krävande för datorns arbetsminne, därför valde författaren 58% av data som icke-testningsdata och inte 95% som det



Figur 5: Cirkeldiagram som illustrerar uppdelningen av talmaterial i mängderna: träning, validering,testning

föreslås i litteraturen (Witten, 2005).

Utan att göra en analys av algoritmernas komplexitet för ANN och inlärningsfunktionerna kan man se, givet mängden data, att träningen kommer att ta tid. Flera författare, bland annat Witten och Frank (Witten och Frank, 2005), påpekar dessutom att artificiella neurala nätverk har en hög inlärningstid, men är snabba att använda när de är färdigtränade. Denna aspekt har påverkat mängden tränings- och valideringsdata i optimeringsfasen, då författaren ville få en tillförlitlig utgångspunkt[11] vid det slutgiltiga valet av akustisk analys (se avsnitt 4.3.2) och det artificiella neurala nätverkets konfiguration (se avsnitt 4.3.3). Träningsmängden reducerades då till 15 minuter med 252 inandningspauser och valideringsmängden behölls som tidigare (18 minuter, 320 inandningspauser), testdata användes inte. Vid reduktion av data såg författaren till att båda talarna i en dialog blev representerade. Efter det slutgiltiga valet användes alla datamängder, i samma proportioner som framgår av Figur 5.

11Med detta menas olika typer av bedömningar utifrån valideringsfasen.

4.3.2 Korpusrepresentation till ANN

Under inlärningsstadiet behöver det artificiella neurala nätverket mönster som består av dels av akustisk information, dels ett facit som talar om vilken kategori mönstret hör till. Under teststadiet saknar nätverket tillgång till facit och ska istället gissa vilken kategori ett tidigare osett mönster hör till utifrån hur det har lärt sig.

Följande stycken beskriver hur författaren har bearbetat datamängderna för träning, validering och test från Kiel Corpus of Spontaneous Speech för att presentera dem för nätverket som en uppsättning akustiska särdrag och rätt kategori per steg i ljudet (det vill säga facit).

Tabell 3: Representation till ANN - den akustiska analysen av talsignalen

Konfiguration	
Typ	**Värde**
Bredd på analysfönser	16 millisekunder
Typ av analysfönster	Hamming
Bredd på steg längs talsignal	8 millisekunder
Frekvensomfång	0 – 8000 Hertz
Antal triangulära Melfilter	20
Antal cepstrala koefficienter	13
Delta koefficienter	13
Delta-delta koefficienter	13

Akustiska särdrag

En representation av det spontana talet i datorn som akustiska sädrag till ANN görs med den sorten av akustisk analys som beskrevs i kapitel tre (se avsnitt 3.4). I Tabell 3 ovan presenteras tekniska detaljer för analysen med grund i inandningspausernas akustiska egenskaper (se avsnitt 2.2) och baserat på utvärdering av HMM/ANN-hybridsystems identifieringsförmåga under valideringsfasen i form av Recall, Precision och F-score.

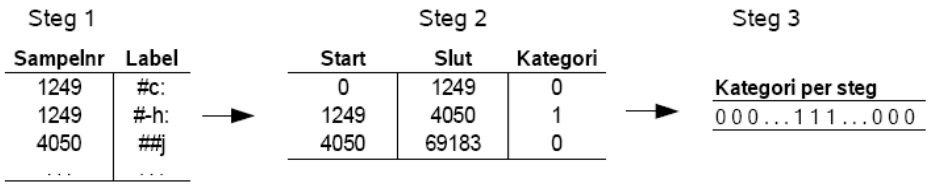
Sammanlagt består representationen av talsignalen av 39 särdrag (13 cepstrala koefficienter med tillhörande 13 delta koefficienter och 13 delta-delta koefficienter) som räknas fram var åttonde millisekund med en akustisk kontext på sexton millisekunder. Representationen blir därmed - 39*antal steg.

Varje sådant steg skall ha en motsvarande kategori, inandningspaus eller icke-inandningspaus. Hur författaren har gått tillväga för att få fram det finns i nästa avsnitt.

Facit

För att få fram facit till nätverket, som består av rätt kategori per steg, används uppmärkningsfilerna från Kiel Corpus of Spontaneous Speech (se avsnitt 4.1.1). Beteckningarna som användes där för att märka upp inandningspauser (och olyckligvis utandningar (se avsnitt 6.2) är: '#h:','#-h:','$h:', '$-h:'.

Som framgår av Figur 6 består omvandlingen av uppmärkningsfilerna, som är länkade med varsin ljudfil, i flera steg. Det första steget är att plocka fram sampelnummer, som står för positionen i ljudet, och motsvarande beteckning. Eftersom författaren måste täcka hela ljudet från 0 till slutet läggs denna information till i steg två. Samtidigt identifieras alla beteckningar som stämmer överens exakt med någon av beteckningarna för inandningspausen, då omvandlas beteckningen till en 1:a som befinner sig spalten för "Kategori". Annars är beteckningen 0, för icke-inandningspaus. Givet denna information produceras en remsa där antingen 1:a eller 0:a upprepas så många gånger som en



Figur 6: Omvandling av uppmärkningsfiler till facit för ANN (Material: G071A008.S1H, Kiel Corpus of Spontaneous Speech, Volym I).

kategori pågår. I Figur 6 pågår kategori 1, inandningspaus, från 1249 till 4050. Om vi omvandlar det till antal steg i ljudet, givet att steget är 8 millisekunder eller 128 sampelnummer (se Tabell 3), får vi cirka 22 ettor i remsan i steg tre (för tekniska detaljer se Bilaga D).

Detta sätt att presentera data reducerar mängden information från

uppmärkningsfilerna precis som den akustiska analysen reducerar mängden information i inspelat tal.

Motivering av akustisk analys

Sammanställning över den valda akustiska analysen finns i Tabell 3, sida 26.

Det framgick av beskrivningen för inandningspausen (se avsnitt 2.2) att intensiteten hos denna är lägre än övrigt tal samt att den verkar föregås och avslutas med att intensiteten sjunker till mycket låg nivå. Av den anledningen valdes 0:te cepstrala koefficienten (se avsnitt 3.4) som ett av särdragen, i förhoppning att i kombination med andra särdrag exempelvis duration och kontext som modelleras av dolda Markov-modeller, ge bra identifikationsresultat. Utan den 0:te cepstrala koefficienten blev resultaten från valideringsfasen sämre i förhållande till när denna var med. Detta är författarens egna bedömning, utan signifikanstest.

Vad gäller frekvensomfånget kunde vi se i Figur 2 (se avsnitt 2.2) att information om ljudtrycksnivåerna vid de olika frekvenserna för två inandningspauser fanns även vid de högre frekvenserna. Det visade sig också vid upprepade valideringar att metodens tillgång till akustiska data som täckte 0 till 8000 Hertz (med och utan delta och delta-delta koefficienter) tydligt förbättrade HMM/ANN-hybridsystems identifieringsförmåga i jämförelse mot frekvensomfånget 0 till 5600 Hertz. Detta är också författarens egna bedömning, utan signifikanstest. Därför valdes frekvensomfånget som framgår av Tabell 3.

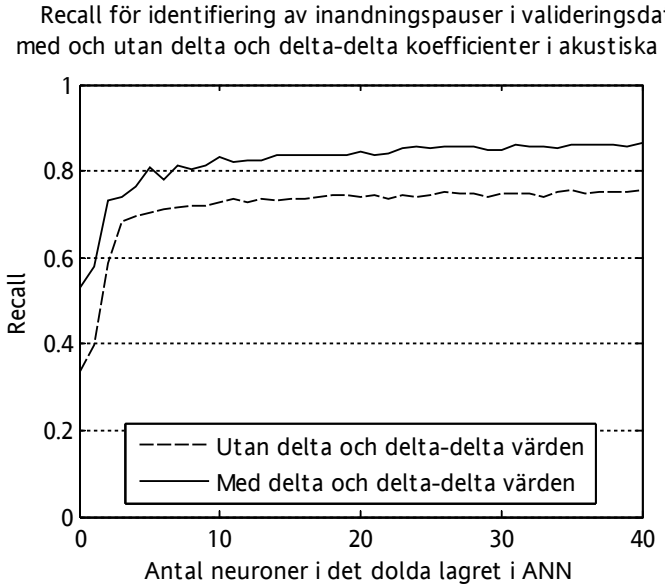
Antal Melfilter valdes till 20 och har inte manipulerats under valideringsstadiet. Antal cepstrala koefficienter hör också till den vanligast använda inom taligenkänning som använder den typen av analys (Elenius och Blomberg, 2000). Författaren provade några gånger med att begränsa antalet till 9, vilket gav sämre resultat enligt samtliga mått än för 13.

Bredden på analysfönstret och steget är också relativt standardiserade (Elenius och Blomberg, 2000). Bredden här är sådan att den fångar tidsrymden bättre än frekvensrymden. Det är möjligt att bredare fönster och därmed bättre upplösning i frekvens skulle ha påverkat metodens identifieringsförmåga. Detta har emellertid inte testats på grund av tekniska skäl. Fönstringsfunktionen som används vid akustisk analys i detta examensarbete är Hamming. Det är en fönstringsfunktion som gynnar de lägre frekvenserna och skapar bra

förutsättningar för vidare analys av talets akustiska egenskaper.

Optimering av akustiska särdrag

Eftersom användning av delta och delta-delta koefficienter gav signifikant olika resultat i Recall (se Figur 7) och Precision (se Figur 8) i sin när- och frånvaro



Figur 7: Recall, median-värde av trettio evalueringar på valideringsdata, med delta och delta-delta koefficienterna i analysen och utan.

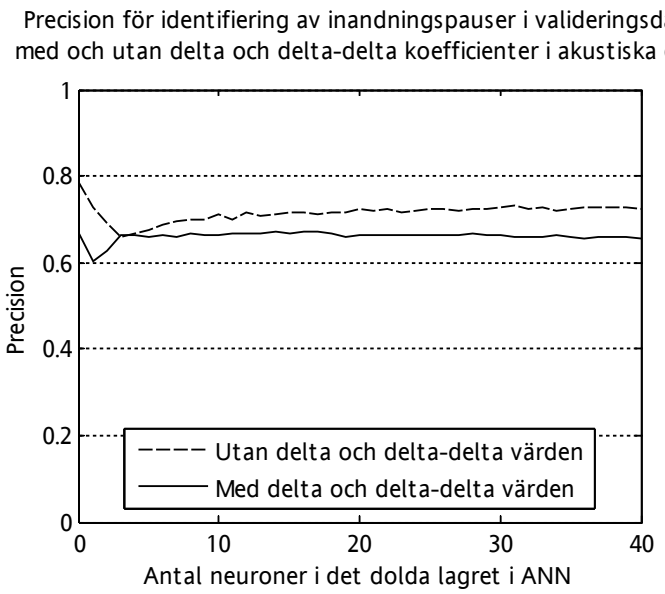
använde författaren skillnaderna i F-score mellan ”med delta och delta-delta värden” och ”utan delta och delta-delta värden” för att välja ut det bättre alternativet. Anledningen till att utvärderingen sker på ordnivå är att utvärdering på stegnivå kan ge en felaktig bild av hur HMM/ANN-hybridsystem identifierar nyckelord (Morgan och Bourlard, 1990). Nedan följer en redogörelse för identifieringsresultat med och utan delta-värden samt signifikanstest för dessa.

Ett icke-parametriskt signifikanstest (Wilcoxon (ranksum i Matlab)) görs, på författarens egna initiativ och i samråd med Joost van de Weijer[12], på data från

12 Personlig korrespondens, 26 oktober 2006.

trettio bedömningar av HMM/ANN-hybridsystems identifieringsförmåga enligt Recall, Precision och F-score, per neuron i det dolda lagret i ANN. Nollhypotesen är att medianerna mellan trettio körningar, per neuron, är lika. Denna avvisas med en signifikansnivå på 0,05. Samma signifikanstest används senare vid val av lägst möjliga antal neuroner i det dolda lagret (se avsnitt 4.3.3).

Enligt signifikanstestet förkastas nollhypotesen för Recall-värden i Figur 7. Det innebär att det finns en signifikant skillnad på om man väljer att ha delta och delta-delta koefficienter eller inte. Utifrån diagrammet är det givet att delta och delta-delta koefficienterna ökar antal identifierade inandningspauser av det antal som faktiskt finns i facit för valideringsdata (Recall).



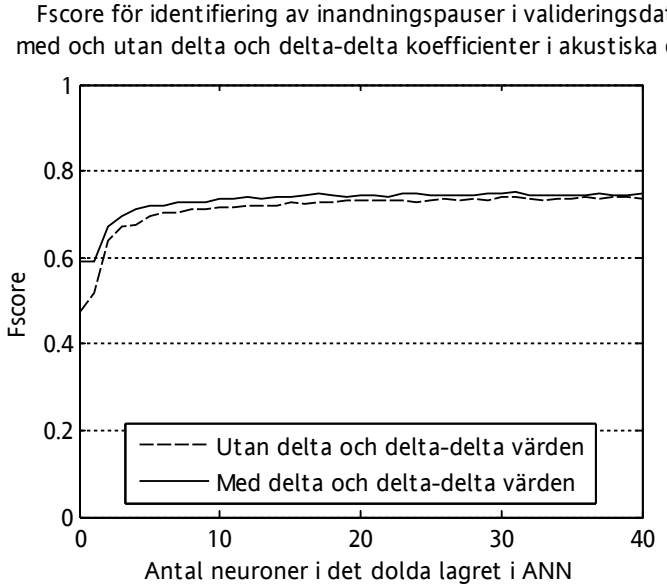
Figur 8: Precision, median-värde av trettio evalueringar på valideringsdata, med delta och delta-delta koefficienterna i analysen och utan.

Men vad gäller Precision, alltså hur många av de identifierade enheterna faktiskt var inandningspauser, är förhållandet omvänt. Akustisk information utan delta och delta-delta värden ger signifikant (Wilcoxon (ranksum i Matlab)) bättre resultat än den med delta-värden. Detta illustreras av Figur 8.

Om man däremot jämför F-score, som uppskattar förhållandet mellan hur mycket metoden träffar i facit och hur många faktiska inandningspauser den identifierar, så ger akustiska data som innehåller delta och delta-delta värden

signifikant bättre resultat, även om det inte riktigt framgår av Figur 9 nedan.

Medel för P-värden, ju lägre desto större chans att nollhypotesen inte är sann, och med andra ord att någon av metoderna är bättre än den andra (Matlab manual), var : F-score – 0,0015; Recall – 0,00000001834; Precision – 0.0294. Eftersom signifikanstest visade via F-score att användning av delta och delta-delta värden var till fördel, så används dessa.



Figur 9: F-score, median-värde av trettio evalueringar på valideringsdata, med delta och delta-delta koefficienterna i analysen och utan.

4.3.3 HMM/ANN-hybridsystem

ANN- arkitektur

Av teoriavsnittet (se avsnitt 3.2) framgick att fler lager än ett, där lager räknas som lager av anslutningar mellan neuroner, gjorde det neurala nätverket kraftfullt. Detta antagande stämmer överens med valideringsresultat för F-score

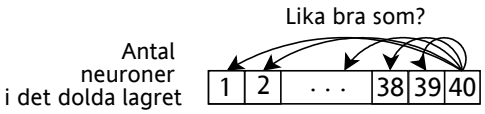
(se Figur 9 sida 31). Figuren visar att noll antal neuroner i det dolda lagret med neuroner ger sämre resultat än fler lager av anslutningar. Av den anledningen bestäms att det artificiella neurala nätverket skall innehålla två lager av anslutningar. Där första lagret anslutningar från inlagret till mellanlagret, och det andra lagret från mellanlagret till utlagret. Inlagret tar in akustiska mönster och utlagret matar ut en kategorisering från det färdigtränade nätverket.

Inlager och utlager

Inlager till nätverket består av lika många neuroner som det finns egenskaper i den akustiska informationen (39 egenskaper; se 4.3.2). Utlagret från nätverket består av en enda neuron som är inställd på att mata ut ett värde mellan 0 och 1. Dessa värden betraktas som sannolikheter, där ”värde>0.5” räknas som en inandningspaus, och ”värde<0.5” räknas som icke-inandningspaus. På så sätt skiljer man mellan dessa två kategorier. Förklaringen till varför författaren har valt att göra på det viset finns i kommande avsnitt.

Mellanlager

Från Figur 7 sida 29, Figur 8 sida 30 och Figur 9 sida 31 (med delta och delta-delta koefficienter), utläser författaren ingen signifikant förbättring efter ett



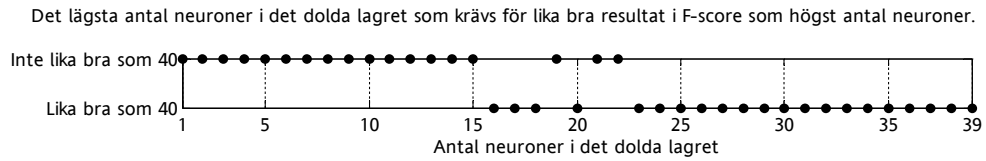
Figur 10: Tillvägagångssätt vid jämförelse av högst antal neuroner med lägre antal neuroner i det dolda lagret.

visst antal neuroner i det dolda lagret. Genom att hålla antalet neuroner i det dolda lagret lågt minskar nätverkets komplexitet och inlärningstid, och som resultat av detta, minnesåtgången vid större mängder träningsdata. Av den anledningen utförs ett icke-parametriskt signifikanstest (Wilcoxon, se Matlab dok.) som ger en fingervisning när ett ökat antal neuroner inte längre förbättrar metodens identifieringsförmåga. På vilket sätt testet används framgår av Figur 10.

Signifikanstestet utgår från utvärderingsresultat, endast i F-score, på akustiska data enligt analysen i Tabell 3, sida. 26. Totalt valideras ett till fyrtio

antal neuroner i det dolda lagret. För varje antal neuroner görs trettio träningar på träningsdata och trettio utvärderingar på valideringsdata. Resultat av jämförelsen finns uppställt i Figur 11.

Anledningen till att författaren tränar trettio gånger för varje antal neuroneristället för att träna en gång och evaluera trettio gånger, är att vissa saker sker slumpmässigt under träningsfasen, och det är det författaren vill komma ifrån genom många träningar. Orsaken till att endast F-score används är att valet av antalet neuroner måste ske enligt en enda parameter. Märk att nollte neuronen, vilket svarar mot inget mellanlager, har valts bort eftersom F-score är lågt i jämförelse mot att ha ett mellanlager (se Figur 9, sida 31). Att antalet körningar är begränsade till trettio är för att författaren ansåg det vara tillräckligt.



Figur 11: Det lägsta antal neuroner i det dolda lagret som krävs för ett lika bra resultat i i F-score som högst antal neuroner (40).

Figur 11 visar att sexton neuroner i det dolda lagret är lika bra som fyrtio neuroner i det dolda lagret. De avvikande antalen är nitton, tjugoett och tjugotvå. Spridningen på F-score för de senare två är lägre än för fyrtio neuroner. Nitton neuroner innehåller däremot flera mycket avvikande värden. Båda företeelserna, tror författaren, beror på slumpen. Därför antas sexton neuroner vara det lägsta möjliga antalet neuroner i det dolda lagret.

Vid tidigare test, hade författaren endast gjort tio träningar och utvärderingar, och baserat sitt val av neuroner i det dolda lagret på det. Det valet var fjorton neuroner i det dolda lagret. Även om det nu, av Figur 11, framgår att sexton neuroner är det lägsta, använder författaren inställningen med fjorton neuroner i det dolda lagret. Detta för att det nätverket är redan färdigtränat på ett större material och utvärderat på testdata. Hur det slutgiltiga nätverket ser ut finns illustrerat i avsnitt 5.1.

Övriga parametrar

Nedan redovisas några tekniska parametrar som saknar presentation i teoriavsnittet, men som ändå kan vara av intresse. Dessa parametrars inverkan på metodens identifieringsförmåga diskuteras inte i kommande avsnitt.

Det artificiella neurala nätverket tränas med flexibel generaliserad deltaregel, det som på engelska heter ”resilient backpropagation”. Denna regel skall vara särskilt bra för mönsterigenkänning (se Matlabs Manual). Felfunktionen, som bestämmer anpassning av vikterna, är den kvadrerade skillnaden mellan önskad output och verklig output (engelska ”mean squared error”). Valideringsdata används för att stoppa inlärningen (engelska ”early stopping”) innan ANN börjar övergeneralisera. All data (akustiska särdrag) normaliseras så att värden har noll i medelvärde och ett i standardavvikelse. All data under träningen bearbetas satsvis (engelska ”batch training”). Inlärningstakten ligger på 0,5 (på en skala mellan 1 och 0). Funktionerna, som nätet strävar efter att anpassa indata mot är: i mellanlagret sigmoid, och utlagret logsig.

HMM/ANN-hybridsystem- arkitektur

Renals och kollegorna (Renals et al., 1994) föreslår att kombinera ANN med dolda Markov-modeller enligt sättet som beskrevs i avsnitt 3.3. Då används Bayes regel på nätverkets a posteori sannolikheter för att få fram emissionssannolikheter. Vi gör på ett lite annorlunda sätt, dock enligt samma princip, för att fram emissionssannolikheterna.

Taligenkänningsproblemet är här reducerat till nyckelordsidentifiering där nyckelordet är en inandningspaus. Samtidigt har vi den omgivningen, icke-inandningspaus, som nyckelordet befinner sig i. Båda alternativen måste täckas av hybrid HMM/ANN-system för att täcka in hur de förhåller sig varandra. Hur en inandningspaus förekommer beror på hur icke-inandningspausen förekommer.

Vi har alltså två alternativ. Det neurala nätverket tränas mot att producera en 1:a om den akustiska informationen i ett steg i ljudet tillhör en inandningspaus, och en 0:a om den akustiska informationen tillhör en icke-inandningspaus. Den funktionen som bestämmer vilka värden som nätverket matar ut ger värden mellan 0 och 1. Dessa värden betraktas vara a posteori sannolikheter.

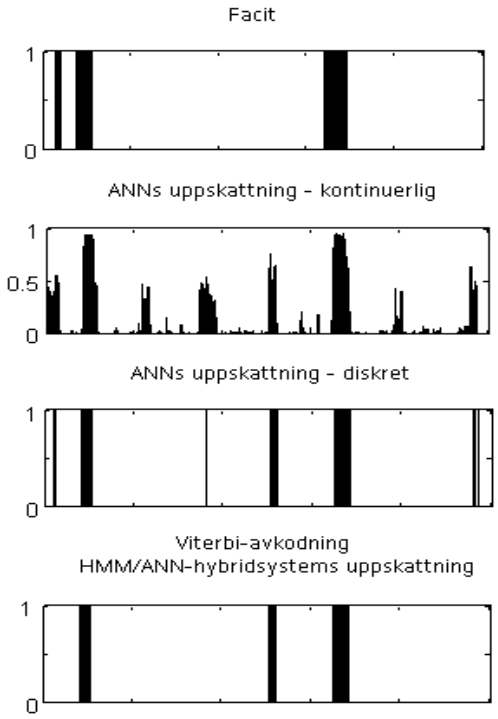
Det nämndes tidigare (se avsnitt 3.3) att dessa a posteori sannolikheter skulle divideras med a priori sannolikheter för tillstånden för att få fram emissionssannolikheter. I fallet här är a priori sannolikheterna för inandningspaus 0,13 och för icke-inandningspaus 0,87. Till en början testade författaren att göra på det här sättet men det visade sig inte gå så bra på grund av hur Viterbi-algoritmen var implementerad i verktygslådan HMMToolbox. Därför tog ett alternativt tillvägagångssätt sin form i samråd med Johan Frid och Anders Sjöström vid ett handledningstillfälle. Det är det tillvägagångsättet som används för att uppskatta både övergångs- och emissionssannolikheter från träningsdata i den dolda Markov-modellen.

Som sagt så matar ANN ut ett värde mellan 0 och 1 för varje steg i det analyserade ljudet. Ett värde som befinner sig under 0,5 betraktas vara en icke-inandningspaus. Ett värde som överstiger 0,5 betraktas vara en inandningspaus. Läsaren kanske tänker varför a priori sannolikheterna inte användes här. Svaret är att den sannolikheten som det neurala nätverket matar ut är en uppskattning av hur stor chansen är att ett steg tillhör en inandningspaus. Och inte en sannolikhet för hur stor chansen är att ett steg tillhör icke-inandningspaus och inandningspaus.

Anledningen till att ett tröskelvärde används är för att omvandla kontinuerliga uppskattningar till diskreta, alltså kategorierna 0 (icke-inandningspaus) och 1 (inandningspaus). Detta steg behövs delvis för att uppskatta hur väl ANN klassificerar på stegnivå, men också för att betrakta denna sekvens som en sekvens av de emitterade symbolerna som modellen genererade när den passerade en sekvens av tillstånd. Givet denna sekvens av symboler kan man nu få fram en matris med emissionssannolikheter för att symbolen '0' eller '1' kommer att genereras när modellen är i tillstånd för 0 – icke-inandningpaus, och samma för tillstånd 1 – inandningspaus. Titta gärna i Figur 13 på sidan 38 och Figur 16 sidan 39. Figur 13 återger helhetsbilden för hur uppskattningen av övergångs- och emissionssannolikheter går till. Figur 16 illustrerar den dolda Markov-modellen. Övergångssannolikheterna bestäms utifrån facit, där tillståndssekvensen är ju känd, för träningsdata.

När HMM/ANN-hybridsystemet skall känna igen nya data så används det neurala nätet, som på vanligt vis genererar en sekvens av a posteori sannolikheter. Dessa diskretiseras med samma tröskelvärde som tidigare (0,5) och sekvensen betraktas vara en sekvens med de emitterade symbolerna för den dolda sekvensen av tillstånd i Markov-modellen– $P(Y|w)$. Givet ANNs

diskretiserade a posteori sekvens uppskattas den mest troliga sekvensen av tillstånd – P(w|Y) via Viterbi algoritmen. Resultatet är den mest sannolika uppskattningen av var inandningspauser och icke-inandningspauser finns. För en grafisk framställning se Figur 12.



Figur 12: Hur uppskattningen av kategorier går till. Diagrammet ska läsas uppifrån och ner. 1 - inandningspaus, 0 - icke-inandningspaus (Exempelfil : G1A007).

4.3.4 Utvärdering

Mönsterklassificering

Den stegvisa utvärderingen är implementerad enligt formlerna i Tabell 2. Facit och hybrid HMM/ANN-systemets uppskattningar jämförs steg för steg.

Identifiering av inandningspauser

Recall, Precision och F-score beräknas enligt formlerna uppställda i avsnitt 3.5. Start och stop positioner från facit jämförs mot start och stop positioner i HMM/ANN-hybridsystems uppskattning. Start och stop positioner markerar när en kategori startar och slutar vilket är likvärdigt med duration.

Om en start och stop position i HMM/ANN-hybridsystems uppskattning svarar mot två eller fler korrekta inandningspauser i facit räknas detta som två eller fler korrekt identifierade inandningspauser, om förhållandet är omvänt räknas fler träffar som en. Antal korrekt identifierade inandningspauser (SP) styrs av krav på överlapp i antal steg som ställs på hybrid HMM/ANN-systems positioner i jämförelse mot facits positioner. Positioner för inandningspauser som inte överlappar är antingen missade (FN) eller felaktiga (FP) inandningspauser.

Baslinje

Baslinjen är implementerad enligt förslag[13] från Anders Sjöström. Baslinjen gissar om ett steg i ljudet hör till en inandningspaus eller någon enhet i utfyllnaden.

Ett ett tal mellan 0 och 1 slumpas fram enligt en likformig fördelning. Detta tal betraktas vara en klassificering av ett steg, i form av en sannolikhet. Stegtillhörigheten bestäms utifrån a priori sannolikheter, beräknade i form av relativ frekvens, för de två olika kategorierna. För kategori inandningspaus är sannolikheten - 0,13, och sannolikheten för kategori utfyllnad – 0,87. Är sannolikheten mindre eller lika med 0,13 har ett steg tillhörande inandningspaus identifierats och en etta läggs till i den remsan som senare jämförs mot facit. Är sannolikheten större eller lika med 0,87 är det istället en nolla som läggs till i samma remsa. Därefter utvärderas gissningen enligt antal korrekt identifierade steg och antal identifierade inandningspauser. Proceduren utförs fem hundra gånger. Ett medelvärde för de olika utvärderingsmåtten räknas därefter fram.

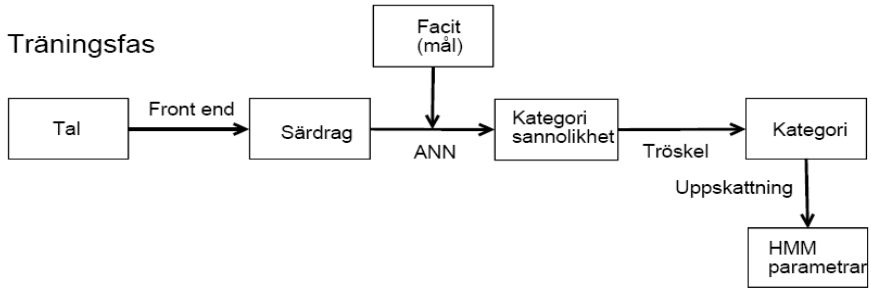
13 Personlig korrespondens, den 24 juli 2006.

5 Resultat

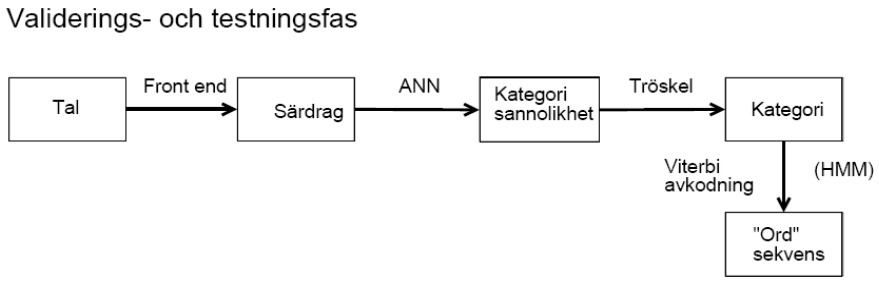
5.1 System

Översikt

Systemöversikten är adapterad från Renals et al. (1994).

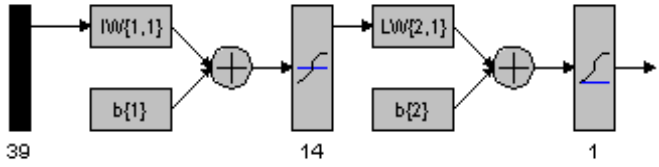


Figur 13: Processerna i systemet under en träningsfas.



Figur 14: Processerna i systemet under validerings- och testningsfas.

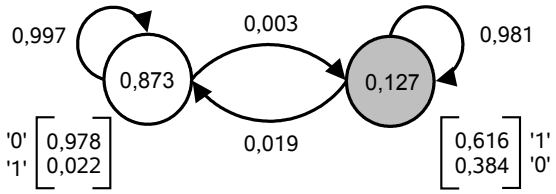
ANN



Figur 15: Det artificiella neurala nätverk som används i HMM/ANN-hybridsystemet för att uppskatta parametrar till den dolda Markov-modellen och för att ge lokala sannolikheter för varje steg i ljudet.

Det artificiella neurala nätverket som är ett resultat från träningsfas (se Figur 13) och som används i valideringss- och testningsfas (se Figur 14).

HMM



Figur 16: Den dolda Markov-modellen med emissionssannolikheterna beräknade med det artificiella neurala nätverket uppställt i Figur 13. Vitt tillstånd är icke-inandningspaus och grått inandnings-paus.

Den dolda Markov-modellen har konfigurerats med hjälp av funktionen *hmmestimate.m* (se Matlab dok). Tillstånds- och övergångssannolikheterna har beräknats utifrån facit i träningsdata, och emissionssannolikheter utifrån utdata från det artificiella neurala nätverket simulerat på träningsdata i Figur 15.

5.2 Utvärdering

Följande resultat fås genom proceduren för test- och valideringsfas från (se Figur 13). Facit till testdata används för att kunna utvärdera systemets identifieringsförmåga. HMM/ANN-hybridsystem modell som används är den som finns uppställd i Figur 16. Testdata (se Figur 4, sida 24) analyseras akustiskt enligt inställningar i Tabell 3, sida 26.

5.2.1 Mönsterklassificering

Resultat av den stegvisa identifieringen, och därmed en bedömning av det neurala nätverkets förmåga att se mönster i nya data, finns uppställda i Tabell 4 och Tabell 5. Utvärderingen utförs enligt tillvägagångssättet beskrivet i 4.3.4. Det artificiella neurala nätverkets konfiguration är enligt Figur 15.

Testdata består totalt av 287 678 steg (cirka 38 minuter). 255 260 (cirka 34 minuter) steg av det totala antalet tillhör icke-inandningspaus, kategori 0. 32 418 (cirka 4 minuter) av det totala antalet tillhör inandningspaus[14], kategori 1.

Tabell 4: Hur ANN:s diskriminerar mellan kategorier på stegnivå

		Identifierad av ANN kategori			
		1		0	
		Mått	Antal steg	Mått	Antal steg
Faktisk kategori	1	SP	20 422	FN	11 996
	0	FP	6 381	SN	248 879

1 – inandningspaus, 0 – icke-inandningspaus (Data: testdata)

Nätverket identifierar korrekt 20 422 av 32 418 steg som inandningspaus. Det identifierar felaktigt 11 996 av 32 418 steg som icke-inandningspaus.

Nätverket identifierar korrekt 248 879 av 255 260 steg som icke-inandningspaus. Det identifierar felaktigt 6 381 av 255 260 steg som inandningspaus.

6 381 (cirka 51 sekunder) steg har alltså identifierats felaktigt bland övrigt tal

14 Och till viss del utandningar, för mer information se 4.1.1 och 6.1.

som inandningspaus. Nätverket har identifierat felaktigt, och därmed missat, 11 996 steg av inandningspauser som icke-inandningspauser.

Bedömning, i procent, av denna diskriminering mellan kategorier på stegnivå jämförs med två andra metoder i Tabell 5, enligt utvärderingsmått beskrivna i avsnitt 3.5, Tabell 2.

Tabell 5: Resultat av stegvis identifiering för ANN, HMM/ANN och Baslinje

Metod	Utvärderingsmått				
Namn	Värde i %				
	SannPositiv	FalskNegativ	SannNegativ	FalskPositiv	Total framgång
ANN	63,0	37,0	97,5	2,5	93,6
HMM/ANN	71,1	28,9	97,2	2,8	94,3
Baslinje	1,7	97,3	98,3	1,7	87,4

(Data: testdata)

Det artificiella neurala nätverket identifierar 63,0% steg som tillhör inandningspauser, och missar 37,0%. Det identifierar 97,5% steg som tillhör icke-inandningspauser, men 2,5% steg identifieras felaktigt som inandningspauser.

HMM/ANN-hybridsystem gör bättre ifrån sig med 8,1% än ANN när det gäller att identifiera steg tillhörande inandningspauser, med 71,1% korrekt identifierade, 28,9% missade. Den är dock sämre än ANN, med 0,3%, på att identifiera steg tillhörande icke-inandningspauser. Antal korrekt identifierade steg tillhörande icke-inandningspauser är 97,2% och antal felaktigt identifierade 2,8%. Högre antal korrekt identifierade steg tillhörande icke-inandningspauser, genom koppling av ANN med HMM och Viterbi-avkodning, fås på bekostnad av fler felaktigt identifierade steg (ANN – 2,5%, HMM – 2,8%) bland icke-inandningspauser.

Baslinjen identifierar endast 1,7% av steg tillhörande inandningspauser, och missar 97,3%. Baslinjen misslyckas, i princip, med att identifiera inandningspauser. Däremot går det riktigt bra, i jämförelse mot de andra två metoderna (0,95% högre i genomsnitt), att identifiera steg tillhörande icke-inandningspauser. Den identifierar 98,3% av steg tillhörande icke-

inandningspauser korrekt och missar endast 1,7%.

Enligt ”Total framgång”, som är den totala bedömningen av det totala antal korrekt identifierade steg, så presterar baslinjen ganska bra, Total framgång – 87,4%. Detta höga värde beror på att steg tillhörande icke-inandningspauser är mer frekventa (88,7% av all testdata) än steg tillhörande inandningspauser. Men både ANN, Total framgår – 93,6%, och HMM/ANN, Total framgång – 94,3%, presterar bättre än baslinjen.

Sammanfattningvis kan författaren konstatera att artificiella neurala nät och HMM/ANN-hybridsystem är bättre på att kategorisera steg än baslinjen, och att HMM/ANN-hybridsystem utgör en förbättring gentemot enbart ANN vad gäller att kategorisera steg.

5.2.2 Identifiering av inandningspauser

Identifieringen görs med tillvägagångsättet beskrivet i 4.3.4 på testdata (38 minuter, 696 inandningspauser), med modellens konfiguration enligt avsnitt 5.1.

Observera att överlappet för HMM/ANN är bestämt till 3 steg, vilket ger optimalt antal identifierade mot facit. Överlappet för baslinjen är 1 steg eftersom baslinjen alltid genererar en gissning per inandningspaus.

Totalt identifierar HMM/ANN-hybridsystem metoden 906 enheter. 632 av dessa är korrekt identifierade inandningspauser, 274 är felaktigt identifierade inandningspauser. Metoden missar 64 inandningspauser. Förhållandet mellan antal korrekta, felaktiga och missade inandningspauser sammanfattas av Tabell 6.

Tabell 6: Antal identifierade inandningspauser för HMM/ANN-hybridsystem och Baslinje

Metod	**Utvärderingsmått**		
Namn	Värde i %		
	Recall	Precision	F-score
HMM/ANN	90,8	66,4	76,7
Baslinje	45,1	6,7	12,1

(Data: testdata)

Tabell 6 visar att HMM/ANN-hybridsystem identifierar 90,8% (632 av 696) av alla inandningspauser. Samtidigt är antal felaktigt identifierade inandningspauser relativt högt (274 enheter av 906), vilket ger 66,4% i Precision. Det procentuella antalet i Recall väger upp det procentuella antalet i Precision vilket ger ett värde på 76,7% i F-score.

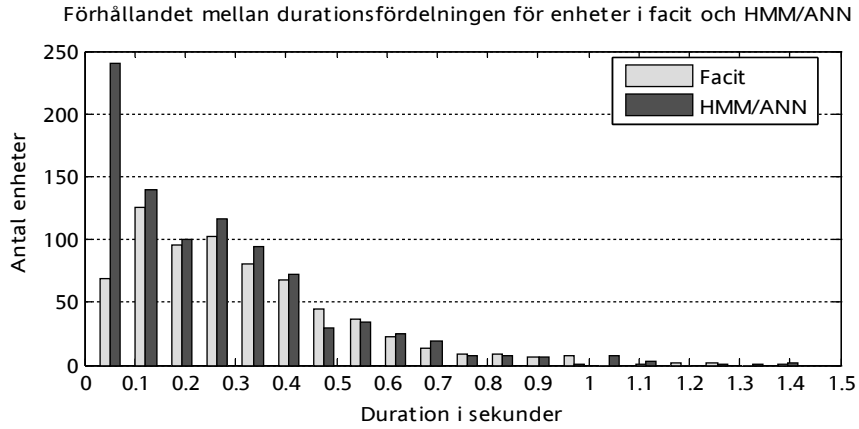
HMM/ANN-hybridsystem ger betydligt bättre resultat än baslinjen enligt samtliga utvärderingsmått. Baslinjen lyckas träffa 45,1% av inandningspauserna i facit, men antal felaktiga enheter är mycket högt, Precision är bara 6,7%. F-score för baslinjen är mycket lägre, 12,1%, än för HMM/ANN, 66,4%.

Sammanfattningsvis kan det konstateras att HMM/ANN-hybridsystemet ger bättre resultat än baslinjen som gissar sig fram. HMM/ANN-hybridsystemet är bra på att träffa enheter som faktiskt är inandningspauser, men detta på bekostnad av flera felaktigt identifierade enheter.

En djupare analys, enligt egenskapen duration av samtliga enheter som identifierats av facit och HMM/ANN-hybridsystem metod i testdata, följer i nästa avsnitt. Förslag och ett mindre test på hur antalet felaktigt identifierade enheter kan reduceras följer i kommande avsnitt (se avsnitt 5.2.3 och 6.3).

5.2.3 Djupare analys

I avsnitt 3.5 beskrevs att en djupare analys av identifiering av inandningspauser kan vara viktig för att förbättra systemets prestanda. I Figur 17) presenteras ett diagram som visar hur många enheter, av samma duration, finns i facit för testdata och i HMM/ANN-hybridsystem:s uppskattning av enheter testdata.



Figur 17: Durationsfördelning för inandningspauser i facit och i HMM/ANN-hybridsystems identifieringar. Baserat på resultat från avsnitt 5.2.2 (Data: testdata).

Av diagrammet framgår att durationsfördelningen mellan enheterna i ”Facit” och i ”HMM/ANN” har liknande mönster. Det som skiljer dessa åt är att betydligt fler enheter identifierade av HMM/ANN-hybridsystem metod har lägre duration än enheterna i facit. Det framgår också av medelvärden, ”HMM/ANN” (0,26 sekunder) är 0,10 sekunder lägre i genomsnitt än i ”Facit” (0,36 sekunder). Den första stapeln i diagrammet visar på den största avvikelsen i antal enheter av viss duration mellan ”HMM/ANN” och ”Facit”, där ”HMM/ANN” innehåller mer än 2/3 delar fler enheter än ”Facit”. Det finns alltså betydligt fler enheter i ”HMM/ANN” som är lägre än 0,1 sekund, än vad det finns i ”Facit”.

Tidigare har nämnts att en eller flera säkerhetsfaktorer är viktiga för optimering av nyckelordsidentifiering. Bättre durationsmodellering från den hybrida HMM/ANN metodens sida skulle kunna vara en sådan säkerthetsfaktor. Mer om detta i avsnitt 6.2.1.

Följande resultat uppnåddes vid tre olika tröskelvärden (*tröskelvärden*

baserade på testdata). Dessa resultat bör betraktas endast som preliminära, hypotetiska resultat. Precision förbättrades med 14,5% för tröskelvärde 0,08 sekunder, som följd 6,8% sämre i Recall, med högre F-score på 5,8%. Baslinjen är inte aktuell här. Vi filtrerar bara bort även de korrekta, därför sänkt Recall. Därför bör man modellera själva fördelnigen.

Tabell 7: Antal identifierade inandningspauser för HMM/ANN-hybridsystem metod efter tillämpning av tröskelvärde för duration.

Tröskelvärde	**Utvärderingsmått**		
Duration i sek.	Värde i %		
	Recall	Precision	F-score
0,07	86,3	78,6	82,3
0,08	84,0	80,9	82,5
0,09	82,3	81,5	81,9

6 Diskussion

6.1 Resultat

Författaren tror att några av de enheterna som HMM/ANN-hybridsystem identifierar felaktigt är exempelvis frikativa brusljud och aspirerade klusiler. HMM/ANN-hybridsystem tar möjligtvis fasta på den lilla biten av en frikativa eller liknande som uppvisar samma ljudenergifördelning som en inandningspaus.

Identifieringsresultat för HMM/ANN-hybridsystem metoden är möjligtvis bättre än mönstermatchningsmetoden utvecklad av Sjöström och Frid (Horne et al., 2005b) men problematiken är densamma. Antal felaktiga enheter är högt på bekostnad av många korrekt identifierade inandningspauser. De felaktiga enheter det handlar om är olika typer av brusljud och brusiga röstkvalitéer, som följer de förutsagda felaktiga enheterna i avsnitt 2.2. De enheter som missas av HMM/ANN-hybridsystem har inte analyserats vilket bör göras för att förbättra framtida identifiering. Möjligen är dessa enheter utandningar.

Det går inte att säga om HMM/ANN-hybridsystem är bättre eller sämre än den mönstermatchande metoden (Horne et. al, 2005b), men eftersom metoden implementerar HMM, som ytterligare ett steg i förbättrad analys av tidsserier, bör det vara bättre. Det är svårt att jämföra med tidigare arbeten eftersom ingen har gjort precis samma sak tidigare med samma metod, och speciellt på samma material. Men av relaterat arbete (Sá et al., 2002) framgick det att inandningar och utandningar gick bra att detektera med hjälp av ANN i ett material som bara innehöll dessa enheter. Eftersom uppgiften att identifiera inandningspauser i spontant kontinuerligt tal är svårare antas resultaten, speciellt efter applicering av en enklare säkerhetsfaktor (se avsnitt 5.2.3), vara relativt bra. Att jämföra mot fonemigenkänningen i kontinuerligt tal i (se avsnitt 2.3) så gick det sämre för HMM/ANN-hybridsystem å andra sidan var den akustiska analysen i det arbetet mycket välavvägd med välmotiverade mått.

Vad gäller systemet i sig så har det implementerats väl både vad gäller metoden men också utvärderingsmetodik. Systemet kan användas precis som det är nu för inlärning och igenkänning av en enhet i tal. Enda kravet är att

utvecklaren måste följa ett visst format (se Figur 6) på uppmärkningsdata för att kunna processas i systemet. Skall systemet utökas till att känna igen fler enheter måste metoden för kopplingen mellan HMM/ANN följas enligt resonemanget i avsnitt 3.3 som finns i Hervé Bourlards och hans kollegors arbeten.

6.2 Data

Korpus

Det har tidigare nämnts att de enheter som de mänskliga segmenterarna i Kiel Corpus har märkt upp också kan innehålla utandningar. När en utandning förekommer, kommer det oftast i samband med att man andas in direkt efteråt, därför är det kanske inte så stort problem egentligen. Tyvärr introducerar utandningen ett mönster som författaren tror gör att antalet falskt positiva är stort, eftersom energin är på väg utåt i dessa ljud.

Akustisk analys

Den akustiska analysen diskuteras ganska uttömmande i avsnitt 4.3.2 om databehandling.

Det jag eftersträvar förutom denna analys är kanske mer intressanta, innovativa tankar, att man väljer en metod som framhäver de egenskaperna hos inandningarna som skiljer de från övrigt tal. Det kräver mer efterforskning och mer medveten akustisk analys.

Det har också funnits tankar om att inkorporera intensitetsutvecklingen över tiden som inandningpauser verkar ha (se Figur 2, sida 10). Författaren tror starkt på att det har gjorts av det HMM/ANN-hybridsystemet eftersom nollte koefficienten, som ju representerar ett stegs intensitet, presenteras tillsammans med dess utveckling över tiden. På så sätt fångas denna utveckling.

Sammanfattningvis kan man säga den nuvarande akustiska analysen är tillfredsställande, med eventuella smärre modifikation kan bli ännu bättre. Däremot så efterfrågas en ny, kunskapsdriven medveten akustisk analys. Där måtten är valda baserade på ”manuella” undersökningar av inandningspauser.

6.3 Metod

Säkerhetsfaktor

Från avsnitt 5.2.3 framgick det att durationsfördelningen för enheterna identifierade av HMM/ANN-hybridsystem var skild från durations-fördelningen på inandningspauserna i facit. Detta innebär att man bör egentligen sträva efter samma fördelning av durationer som det är i facit. Det finns också en artikel som kritiserar durationsmodelleringen i HMM, och de ger förslag på hur man skulle kunna modellera durationer explicit (Tóth och Koscor, 2005).

Det konstateras i djupare analys att medelvärdet än betydligt lägre än medelduration för inandningspauser föreslagna av Conrad och Schönle (se avsnitt 2.2). Det är alltså många korta enheter i både facit och HMM/ANN. Författaren tror att det till största del beror på valet av träningsdata. Som det nämndes är enheter som kan ställa till det är ordinterna in- och utandningar samt utandningar. Det bör också poängteras att duration på inandningspauserna mycket väl kan bero på taltempo.

Fysiologiska mätningar

I introduktionskapitlet kunde vi se (se Figur 1, sida 2) att mätningar av bröstkorgens omkrets (Conrad och Schönle, 1979, sida 254), också kallat för pneumografi, under spontan monolog, markerade inandningspausernas närvaro tydligt. De motsvarades av en hastig och hög ökning av bröstkorgens omkrets i förhållande till övriga mätningar av bröstkorgens omkrets. Det bästa hade varit om man alltid kunde använda den metoden för att identifiera inandningspauser. Kruxet är emellertid att metoden kräver att den person, vars tal man undersöker, är fysiskt närvarande och har en speciell anordning kring sin bröstkorg.

Om man ska använda identifiering av inandningspauser som en komponent i ett automatiskt talparsingssystem, möts man av det faktum att endast talsignalen är tillgänglig. Trots detta kan man indirekt använda pneumografi vid automatisk igenkänning av inandningspauser i tal. Genom en synkroniserad inspelning av pneumografiska mätningar och talsignalen kan man få fram de positioner i talet som motsvarar inandningspauser. Dessa data kan användas som träningsdata för en maskininlärningsmetod. Att få fram mycket talmaterial, från

olika personer, miljöer, talsituationer, där inandningspauserna är uppmärkta med hög korrekthet, skulle gå mycket snabbare relativt det mödosamma arbetet med manuell uppmärkning. Denna ansats förändrar inte behovet av en välavvägd akustisk analys.

Pneumografi kan däremot användas rakt av vid arbetet med den bakomliggande hypotesen om tidsbegränsningar på prosodiska fraser i spontant tal. Mer om hur i nästa avsnitt.

6.4 Hypotes

Om man vill undersöka hypotesens antaganden i (Horne et al., 2005a). Så bör man använda pneumografi eftersom det verkar vara den mest tillförlitliga metoden för tillfället för att identifiera inandningspauser.

För att kunna göra relevanta jämförelser och tester får man utgå från en eller flera personer som man studerar i olika talsituationer. Denna studie måste inkludera en undersökning av hur gränsmarkörerna mellan talproduktionsenheterna ser ut och en statistisk beskrivning av deras förekomster och natur.

Man bör också ta kontakt med de som har gjort studier om inandningspausens roll i tal, bland annat Winkworth, Hird och Kirsner, och Conrad och Schönle, och återanvända deras arbeten på ett mycket aktivt och medvetet sätt eftersom mycket information finns att hämta däri.

7 Slutsats

Givet svårighetsgraden på uppgiften och metoden som har använts, och i relation till tidigare arbeten är slutsatsen att det gick bra att identifiera inandningspauser i spontant kontinuerligt tal med många talare. Det finns emellertid utrymme för förbättring genom introduktion av ytterligare metoder, vilket är samstämmigt med Hornes, Frids och Sjöströms slutsats (se Horne et al., 2005b).

Författaren går ett steg längre och ger konkreta exempel på metoder för att förbättra identifiering av inandningspauser. Ett av förslagen är att som försteg identifiera de enheter som ofta är förväxlas med inandningspauser, exempelvis frikativor och klusiler. Förslagen är också att använda en mer sofistikerad akustisk analys (Abdelatty Ali et al., 1998), inklusive förbättrad modellering av duration (László och András, 2005), samt en högkvalitativ korpus. Författaren ger ett konkret exempel på hur ett sådant korpus kan skapas.

Litteraturförteckning

Abdelatty Ali, A.M., Van der Spiegel, J, Mueller, P. (1998). "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants", ICASSP '98. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, 961-964.

Conrad, B., Schönle, P. (1979). ”Speech and respiration”, European Archives of Psychiatry and Clinical Neuroscience volume 226, number 4, 251-268, Springer.

Elenius, K. och Blomberg, M. (2000). ”Automatisk igenkänning av tal”, Institutionen för tal, musik och hörsel, Kungliga Tekniska Högskolan.

Englund, C. (2004). "Speech recognition in the JAS 39 Gripen aircraft - adaptation to speech at different G-loads". Master Thesis in Speech Technology, supervisor and examiner: Kjell Elenius. Department of Speech, Music and Hearing, Royal Institute of Technology.

Gold, B., Morgan, N. (2000). ”Speech and signal processing: processing and perception of speech, and music”, ISBN 0471351547, John Wiley & Sons, Inc.

Gordon, M., Barthmaier, P., Sands, K.(2002)."A cross-linguistic acoustic study of voiceless fricatives", Journal of the international phonetic association, number 2, volume 32, 141-174. Cambridge University Press.

Henderson, A., Goldman-Eisler, F. & Skarbek, A. (1965). ”Temporal patterns of cognitive activity and breath control in speech”, Language and Speech, 8, 236-242.

Hird, K. and Kirsner, K. (2002). “The relationship between prosody and breathing in spontaneous discourse”, Brain and Language, 80, 536-555.

Hixon, T. J. (1973). ”Respiratory function in speech”, Normal Aspects of Speech, Hearing and Language, 73-125, (edited by Minifie F.D., Hixon T. J., Williams F.), Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Holmes, J., Holmes, W. (2001). ”Speech synthesis and recognition, 2:nd edition”, ISBN 0748408576, Taylor & Francis.

Horne, M., Frid, J. and Roll M. (2005a). “Timing restrictions on prosodic phrasing”, Nordic Prosody IX. Frankfurt am Main: P. Lang.

Horne, M., Frid, J. and Johansson, A. (2005b). “Using cepstral coefficients for inhalation pause detection in spontaneous speech”, Proceedings of 10th International Conference on Speech and Computer (SPECOM 2005).

Iwarsson, J. (2000). "Perceptual detection of inhalations in reading" (abstract), Speech, Music and Hearing Quarterly Progress and Status Report, volume 41 (4).

Jessen, M. (1999) “Redundant aspiration in German is primarily controlled by closure duration”, Proceedings XIVth ICPhS, volume 2, 993-996.

Kohler, K. J., Pätzold M., Simpson A.P. (1995). “From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech”, Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel, 29.

Matlab Manual [www] (2006-11-27) http://www.mathworks.com/access/helpdesk/help/techdoc/matlab.shtml

McFarland, D. H. (2001). “Respiratory Markers of Conversational Interaction”, Journal of Speech, Language, and Hearing Research, 44, 128-143.

McCowan I., Moore D., Dines J., Gatica-Perez D., Flynn M., Wellner P., Bourlard H. (2005). "On the use of information retrieval measures for speech recognition evaluation", IDIAP research rapport 04-73.

Morgan, N., Bourlard, H. (1990). "Continuous speech recognition using multilayer perceptrons with hidden Markov models", International Conference on Acoustics, Speech, and Signal Processing, 1990, ICASSP '90., volume 1, 413-416.

Renals, S., Morgan, N., Bourlard, H., Cohen, M., Franco, H. (1994). "Connectionist Probability Estimators in HMM Speech Recognition", IEEE Transactions on Speech and Audio Processing, number 1, volume 2, part 2, 161-174.

Sá, R.C., Verbandt, Y. (2002). "Automated breath detection on long-duration signals using feedforward backpropagation artificial neural networks", IEEE Transactions on Biomedical Engineering, number 10, volume 49, 1130-1141.

Schultz T., Rogina I. (1995)."Acoustic and language modeling of human and nonhuman noises for human-to-human spontaneous speech recognition", ICASSP '95, International Conference on Acoustics, Speech, and Signal Processing, 1995, volume 1, 293-296.

Sundaram, S., Narayanan, S. (2003). ”An empirical test transformation method for spontaneous speech synthesizers”, Eurospeech 2003.

Szöke I., Schwarz P., Matejka P., Burget L., Fapso M., Karafiát M., Cernocký J. (2005)."Comparison of Keyword Spotting Approaches for Informal Continuous Speech", Proceedings of 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms.

Tóth, L. och Koscor, A. (2005). "Explicit Duration Modelling in HMM/ANN Hybrids", Lecture Notes in Computer Science, volume 3658, 310–317.

The Kiel corpus of spontaneous speech, volume I (1995). [cd-rom]. Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universität zu Kiel.

The Kiel corpus of spontaneous speech, volume II (1996). [cd-rom]. Institut für Phonetik und digitale Sprachverarbeitung der Christian-Albrechts-Universität zu Kiel.

Ward, W. (1989). "Modelling Non-Verbal Sounds for Speech Recognition", Proc. DARPA Workshop on Speech and Natural Language, 47-50.

Weilhammer, K., Schiel, F. (1999). “Investigations of language structure by means of language models incorporating breathing and articulatory noise”, International Conference of Phonetic Sciences.

Wightman C.W., Ostendorf M. (1994). "Automatic labeling of prosodic patterns", IEEE Transactions on Speech and Audio Processing, number 4, volume 2, 469-481.

Wightman C.W., Ostendorf M. (1991). "Automatic recognition of prosodic phrases", ICASSP '91. 1991 International Conference on Acoustics, Speech, and Signal Processing, volume 1, 321-324.

Winkworth, A. L., Davis, P. J., Adams, R. D., & Ellis, E. (1995). “Breathing patterns during spontaneous speech”, Journal of Speech and Hearing Research, 38, 124-144.

Witten, I. H.. Frank, E. (2005). “Data mining: practical machine learning tools and techniques”, ISBN 0120884070, Elsevier Inc.

A Uppmärkningsfil från korpus

Exempel på en fil ur Kiel Corpus of Spontaneous Speech, volume 2 (Källa: README_VOL2.1ST)

”

```
g101a004.s1h
UTB004: <A> ja , das ist mir sehr recht <A> .
oend
h:  j 'a:  ,  d a s+  Q I s t+  m i: 6+  z 'e: 6  r 'E C t
h:.

kend
c:  h:  j 'a:  ,  d a s+  Q- -q I s t-+  m i:6+  z 'e:6
-:k r 'E C t -h  h:  .

hend
1 #c:
 1 #h:
4261 ##j
5273 $'a:
6351 #,
6351 ##d
.   .
.   .
.   .
```

”

B Information om talarna i korpus

Information om talarna i ljudfilerna som har använts som tränings-, validerings- och testningsdata i examensarbetet ”Automatisk identifiering av Inandningspauser i Spontant Tal”. Informationen är adapterad från filerna README_VOL1.1ST, volym 1 Kiel Corpus Of Spontaneous Speech och README_VOL2.1ST, volym 2 Från Kiel Corpus of Spontaneous Speech .

Förklaring till akronymerna i kolumn ”Dialekt”
Talad tyska (dialekttillhörigheten bestämts av talarna själva):

NI - Niedersachsen
SH - Schleswig-Holstein
HH - Hamburg
ND - North German
RP - Rheinland-Pfalz
BE – Berlin
NW - North Rhine-Westfalia.

Träningsdata
Kiel Corpus of Spontaneous Speech, volume 2

Dialog Id	Dialog Tillgängliga mappar	Kön	Födelsedatum	Dialekt
G10	1A - 7A	f	06.08.68	SH
		f	20.05.68	NW
G11	1A - 7A	m	06.03.72	NI
		m	04.08.68	SH
G12	1A - 7A	m	20.08.59	SH
		f	18.05.66	NI
G14	2A - 7A	m	24.09.69	SH
		m	28.07.70	SH
G19	1A, 3A - 7A	f	05.07.72	ND
		f	29.05.74	ND
G21	1A – 2A, 4A - 7A	m	03.12.68	ND
		m	14.07.68	SH

Valideringsdata
Kiel Corpus of Spontaneous Speech, volume 1 & 2

Dialog ID Dialog	Tillgängliga mappar	Kön	Födelsedatum	Dialekt
G20	2A	m	10.04.67	SH
		m	27.05.67	SH
G25	1A – 5A, 7A	f	.73	BE
		f	.73	ND
G27	4A	f	20.01.69	ND
		f	04.03.70	SH
G28	7A	m	25.06.57	RP
		m	07.05.68	ND
G29	7A	m	05.05.65	HH
		m	11.08.59	HH
G30	6A	f	15.07.72	HH
		f	25.01.73	HH
G31	1A – 4A, 6A - 7A	m	06.04.74	SH
		m	26.02.74	SH

Testningsdata
Kiel Corpus of Spontaneous Speech, volume 1

Dialog ID Dialog	Tillgängliga mappar	Kön	Födelsedatum	Dialekt
G07	1A - 7A	m	02.10.69	SH
		m	12.12.67	NI
G08	1A – 7A	m	03.06.67	SH
		m	17.06.68	SH
G09	1A - 7A	f	19.12.68	NI
		f	19.07.65	NI

C Lista över använda filer från korpus

Lista över filer från Kiel Corpus of Spontaneous Speech volymerna 1 och 2 under respektive datamängd. Uppdelning enligt Figur 4 och 5, sidorna 24 och 25.

Träningsdata (258 filer)

Dialog ID

G10	G11	G12	G14	G19	G21
G101A001	G111A000	G121A000	G142A000	G191A000	G211A000
G101A003	G111A002	G121A002	G142A002	G191A002	G211A003
G101A006	G111A004	G121A004	G142A005	G191A005	G211A005
G101A008	G111A006	G121A006	G142A009	G191A008	G211A007
G101A010	G111A009	G121A008	G143A000	G191A010	G211A009
G101A012	G111A011	G121A011	G143A003	G191A014	G211A012
G101A014	G111A014	G121A013	G143A005	G191A018	G211A014
G101A016	G112A000	G121A015	G143A007	G191A020	G212A000
G101A018	G112A002	G121A017	G144A000	G191A026	G212A002
G102A002	G112A005	G121A020	G144A003	G191A028	G212A005
G102A004	G112A010	G121A023	G144A005	G191A030	G212A008
G102A006	G112A012	G121A026	G145A000	G191A032	G212A010
G102A008	G113A000	G122A001	G145A002	G191A036	G214A001
G102A011	G113A003	G122A003	G145A005	G191A038	G214A003
G102A014	G113A006	G122A005	G145A007	G193A000	G214A007
G102A016	G113A008	G122A007	G145A009	G193A002	G214A009
G103A000	G113A010	G122A010	G145A011	G193A005	G215A000
G103A002	G113A012	G122A012	G146A000	G193A009	G215A002
G103A005	G114A000	G122A014	G146A002	G194A000	G215A004
G103A007	G114A002	G123A000	G146A004	G194A002	G215A006
G103A009	G114A004	G123A003	G146A007	G194A005	G215A010
G103A013	G114A006	G123A008	G147A000	G194A010	G215A012
G103A015	G114A008	G123A010	G147A003	G194A012	G216A001
G103A017	G114A010	G123A012	G147A005	G195A000	G216A003
G103A019	G114A012	G123A015		G195A002	G216A006
G104A001	G114A014	G123A017		G195A004	G216A008
G104A003	G115A000	G124A000		G195A008	G216A010
G104A007	G115A002	G124A002		G195A011	G216A014
G104A010	G115A005	G124A005		G195A013	G217A001
G104A012	G115A007	G124A008		G195A018	G217A003
G104A015	G115A010	G124A011		G196A000	G217A005
G104A017	G115A012	G124A013		G196A003	G217A007
G105A000	G115A014	G124A016		G196A006	G217A009
G105A002	G115A016	G124A018		G196A009	
G105A005	G115A018	G125A000		G196A013	
G105A007	G116A001	G125A002		G196A016	
G105A009	G116A003	G125A004		G196A019	
G105A012	G116A005	G125A006		G196A022	

Träningsdata (258 filer)

G105A015	G116A007	G125A008	G197A001
G105A017	G116A009	G125A010	G197A007
G105A019	G117A001	G125A012	G197A011
G106A000	G117A003	G125A014	
G106A002	G117A005	G125A017	
G106A004	G117A007	G125A019	
G106A007	G117A009	G125A024	
G106A009	G117A011	G126A001	
G106A011	G117A013	G126A005	
G106A015		G126A008	
G106A020		G126A010	
G107A000		G126A014	
G107A003		G126A016	
G107A006		G126A018	
G107A010		G126A022	
G107A012		G127A000	
G107A015		G127A003	
		G127A006	
		G127A008	
		G127A010	

Valideringsdata (147 filer)

			Dialog ID			
G20	G25	G27	G28	G29	G30	G31
G202A000	G251A000	G274A000	G287A000	G297A000	G306A000	G311A003
G202A001	G251A001	G274A001	G287A001	G297A001	G306A001	G311A005
G202A002	G251A003	G274A002	G287A002	G297A002	G306A003	G311A007
G202A003	G251A004	G274A003	G287A005	G297A003	G306A004	G312A000
G202A004	G251A005	G274A006	G287A007	G297A004	G306A006	G312A002
G202A007	G251A007	G274A008	G287A010	G297A005	G306A013	G312A007
G202A008	G251A009	G274A009	G287A012	G297A006	G306A017	G312A011
G202A009	G251A010	G274A010		G297A007	G306A019	G312A012
	G251A011				G306A021	G312A013
	G251A012				G306A023	G312A019
	G251A013					G313A002
	G251A014					G313A003
	G251A015					G313A004
	G251A016					G313A005
	G251A017					G313A009
	G251A019					G314A000
	G251A020					G314A004
	G251A021					G314A008
	G251A022					G314A012
	G251A023					G314A013
	G251A025					G314A014
	G251A027					G314A015
	G251A028					G316A001

Valideringsdata (147 filer)

G251A029
G251A031
G251A032
G251A033
G252A000
G252A001
G252A002
G252A003
G252A005
G252A006
G252A007
G252A008
G252A009
G252A010
G252A011
G253A000
G253A001
G253A003
G253A004
G253A005
G253A006
G254A000
G254A001
G254A002
G254A004
G254A005
G254A006
G254A008
G254A010
G254A011
G254A012
G255A000
G255A001
G255A003
G255A004
G255A005
G255A006
G255A007
G255A008
G255A009
G255A010
G255A011
G255A012
G255A013
G257A000
G257A001
G257A003
G257A005
G257A007
G316A002
G316A005
G316A007
G316A008
G316A010
G317A000
G317A001
G317A007
G317A009
G317A010

Valideringsdata (147 filer)

G257A008

Testningsdata (318 filer)

Dialog ID

G07	G07	G08	G08	G09	G09
G071A000	G074A006	G081A000	G084A007	G091A000	G094A012
G071A001	G074A007	G081A001	G084A008	G091A001	G094A013
G071A002	G074A008	G081A002	G084A009	G091A002	G094A014
G071A003	G074A009	G081A003	G084A010	G091A003	G094A015
G071A004	G074A010	G081A004	G084A011	G091A004	G094A016
G071A005	G074A011	G081A005	G085A000	G091A005	G094A018
G071A007	G074A012	G081A006	G085A001	G091A006	G094A020
G071A008	G074A014	G081A008	G085A002	G091A007	G094A021
G071A009	G074A015	G081A009	G085A003	G091A008	G094A022
G071A010	G074A016	G081A010	G085A004	G091A010	G094A023
G071A011	G074A017	G081A011	G085A005	G091A011	G094A024
G071A014	G074A018	G081A012	G085A006	G091A012	G094A026
G071A015	G075A000	G081A013	G085A007	G091A013	G094A027
G071A016	G075A001	G081A014	G085A008	G091A018	G094A028
G071A019	G075A002	G081A015	G085A009	G091A019	G094A029
G071A020	G075A003	G081A017	G085A010	G091A020	G094A030
G072A000	G075A004	G081A018	G085A011	G091A021	G095A000
G072A001	G075A005	G082A000	G085A012	G091A023	G095A001
G072A003	G075A006	G082A001	G085A013	G091A024	G095A004
G072A005	G075A008	G082A002	G085A014	G091A025	G095A005
G072A006	G075A009	G082A003	G085A015	G091A027	G095A006
G072A007	G075A010	G082A004	G085A016	G091A028	G095A007
G072A008	G075A011	G082A005	G086A000	G091A029	G095A008
G072A009	G075A012	G082A006	G086A001	G091A031	G095A009
G072A010	G075A013	G082A007	G086A002	G091A033	G095A010
G072A011	G076A000	G082A008	G086A003	G091A034	G095A011
G072A012	G076A001	G082A009	G086A004	G092A000	G095A012
G072A014	G076A002	G082A010	G086A005	G092A001	G095A013
G072A015	G076A003	G082A011	G086A006	G092A002	G095A014
G072A017	G076A004	G082A012	G086A007	G092A004	G095A016
G072A018	G076A005	G082A013	G086A008	G092A006	G096A000
G072A019	G076A006	G082A014	G086A010	G092A008	G096A001
G072A020	G076A007	G083A000	G086A012	G092A009	G096A002
G073A000	G076A008	G083A001	G086A013	G092A012	G096A003
G073A001	G076A009	G083A002	G086A014	G092A014	G096A004
G073A002	G076A010	G083A003	G086A015	G092A016	G096A005
G073A003	G076A011	G083A004	G086A016	G092A018	G096A006
G073A004	G076A012	G083A005	G086A017	G092A020	G096A007
G073A005	G076A013	G083A007	G086A018	G092A021	G096A008
G073A006	G076A014	G083A008	G086A019	G092A022	G096A009
G073A008	G076A015	G083A009	G086A020	G092A024	G096A010
G073A009	G077A000	G083A010	G087A000	G092A026	G096A011
G073A010	G077A001	G083A011	G087A001	G092A027	G096A012

Testningsdata (318 filer)

G073A011	G077A002	G083A012	G087A002	G093A000	G096A014
G074A000	G077A003	G083A013	G087A003	G093A001	G096A015
G074A002	G077A004	G084A000	G087A004	G093A002	G096A016
G074A003	G077A005	G084A001	G087A005	G093A003	G096A017
G074A004	G077A006	G084A002	G087A006	G093A004	G096A018
G074A005	G077A007	G084A003	G087A007	G093A006	G096A019
		G084A004	G087A008	G093A008	G097A000
		G084A006		G093A010	G097A001
				G093A012	G097A003
				G093A013	G097A004
				G093A014	G097A005
				G094A000	G097A006
				G094A001	G097A007
				G094A002	G097A008
				G094A004	
				G094A006	
				G094A008	
				G094A009	
				G094A010	