Ofelia – en ny syntesröst

En studie om talsyntes i allmänhet och konkateneringssyntes i synnerhet

Adina Svensson

D-uppsats i datalingvistik
Handledare: Johan Frid
Institutionen för Lingvistik
Lunds universitet

2001-12-10

Sammandrag

I denna uppsats ges en allmän beskrivning av olika sätt att skapa talsyntes samt en mer noggrann beskrivning av hur man på relativt kort tid kan skapa en ny syntesröst med konkateneringssyntes. Den nya syntesrösten är unik eftersom det tidigare inte skapats någon kvinnlig syntesröst med sydsvensk dialekt. I inledningen finns en översikt över talsyntesens historia, beskrivning av tre olika sorters talsyntes samt en kort beskrivning av text-till-tal-syntes. Sedan beskrivs arbetet med pilotprojektet till den nya syntesrösten som var till för att utvärdera inspelningsmiljö, talarens röst och de metoder som användes vid inspelning, segmentering och vågformsgenerering. Resultaten visar bl.a. att det är stor skillnad i röstkvalitet då vågformen genereras på två olika sätt. I beskrivningen av huvudprojektet visas skillnader mellan att skapa en pilotröst med ett litet antal talljud och att skapa en syntesröst som ska klara av ett helt språk. Här måste man t.ex. ta ställning till vilka ljud som ingår i den dialekt som syntesrösten är avsedd att ha samt om den bör kunna uttala några utländska ord innehållande talljud som inte finns i grundspråket. I huvudprojektet gjordes syntesen mer fullständig än i pilotprojektet genom att bokstav-till-ljud-regler skapades så att det blev möjligt att skriva in text för omvandling till tal till skillnad från pilotprojektet där man var tvungen att skriva in fonetiska tecken.

Innehållsförteckning

1. Inledning

1.1 Syfte

Syftet med denna uppsats är att beskriva olika sätt att skapa talsyntes samt att redogöra för tillvägagångssätt, problem och resultat i arbetet med att skapa en ny, kvinnlig syntesröst med konkateneringssyntes. Konkateneringssyntes bygger på att man sätter ihop förinspelade talljudsenheter som ofta har lagrats i en databas. En utförligare förklaring finns i avsnittet "Konkateneringssyntes" nedan. Anledningen till att jag valt att skapa en ny syntesröst är dels att jag vill lära mig hur det går till men framför allt att det inte tidigare gjorts någon kvinnlig syntesröst med sydsvensk dialekt. MBROLA-projektet, som samlar och sprider talsynteser för olika språk, har tidigare inte haft någon fri, kvinnlig, svensk syntesröst. Arbetet med den nya syntesrösten fokuseras på framställningen av en difondatabas. Begreppet difondatabas förklaras i stycket "Konkateneringssyntes" nedan. Huvudsyftet är alltså inte att skapa en fullständig text-till-tal-syntes. Begreppet text-till-tal-syntes förklaras i avsnittet "Text-till-tal-syntes" nedan.

1.2 Teoretisk bakgrund

I detta avsnitt kommer jag först att beskriva viktiga händelser i talsyntesens utveckling från de första helt mekaniska maskinerna till dagens datoriserade talgeneratorer. Därefter kommer en kort beskrivning av begreppet text-till-tal-syntes och sist en beskrivning av tre olika sorters talsynteser; artikulatorisk syntes, formantsyntes och konkateneringssyntes. I avsnittet "Historik" har jag till stor del hämtat data från Lemmetty (1999). Där informationen hämtats från andra källor anges dessa direkt i texten.

1.3.1 Historik

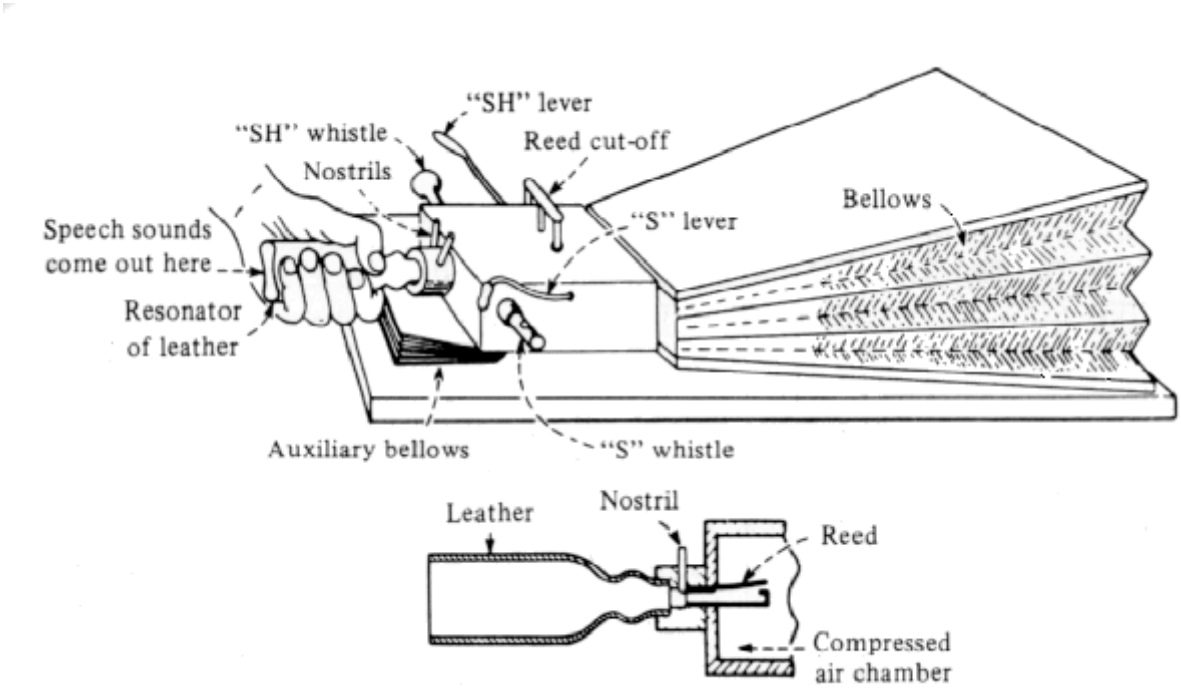
Under andra halvan av 1700-talet gjordes de första försöken att syntetisera mänskligt tal. Ch. G. Kratzenstein lyckades då producera vokaler genom att skapa akustiska resonatorer som liknade det mänskliga ansatsröret och aktivera dem med vibrerande rörblad, liknande de som finns i vissa musikinstrument, t.ex. klarinett och saxofon.

I slutet av 1700-talet uppfann Wolfgang von Kempelen den första talande maskinen som var helt mekanisk. Den hade en tryckkammare som lungor, ett vibrerande rör som stämband och ett läderrör som ansatsrör. Genom att ändra formen på läderröret kunde han producera olika vokalljud. Konsonanter åstadkoms genom att luften fick försvinna ut genom 4 olika förträngda passager som reglerades för hand. Maskinen kunde producera enstaka ljud och några

ljudkombinationer. En rekonstruktion av maskinen byggdes på 1800-talet av Sir Charles Wheatstone utifrån von Kempelens beskrivning. Wheatstones maskin visas i Figur 1.

1835 konstruerade Joseph Faber "Euphonia", en maskin som kunde producera vanligt tal, viskat tal och sång. Euphonia hade en modell av tungan och en faryngal hålighet vars form kunde kontrolleras. Blåsbälgen sköttes via en pedal, medan resten kontrollerades via ett tangentbord. (Traunmüller 1997-2000)

På 1930-talet utvecklade Homer Dudley den första elektroniska talgeneratorn, VODER (Voice Operating Demonstrator). Den var inspirerad av VOCODER (Voice Coder) som hade utvecklats av Bell Laboratories några år tidigare. VOCODER skulle analysera tal och få fram akustiska parametrar som kunde användas av en talgenerator för att rekonstruera den ursprungliga talsignalen. VODER bestod av en handspak med vilken man kunde välja tonande eller tonlös ljudkälla och en fotpedal för att kontrollera grundtonsfrekvensen. Källsignalen



Figur 1. Wheatstones rekonstruktion av von Kempelens talande maskin. Bilden är hämtad från http://www.haskins.yale.edu/haskins/HEADS/SIMULACRA/kempelen.html och kommer ursprungligen från James L. Flanagan, "Speech Analysis, Synthesis and Perception", Springer-Verlag, 1965.

leddes genom tio bandpassfilter, där utsignalens styrka styrdes för hand. Det krävdes en hel del kunskap för att spela[1] en mening på VODER och röstkvaliteten var långt ifrån bra. Idén och strukturen bakom VODER liknar den som senare har använts vid s.k. källa-filter-baserade[2] talsynteser.

1951 konstruerade Franklin Cooper m.fl. en talgenerator som kallades Pattern Playback. Den fungerade som en spektrograf fast baklänges, dvs. den läste spektrogram och framställde motsvarande ljud. Spektrogrammen kunde vara konstruerade utifrån tal eller ritade av användaren för att testa vilka ljud som motsvarade olika akustiska korrelat. 1953 introducerade Walter Lawrence den första formantsyntesen, PAT (Parametric Artificial Talker). PAT bestod av tre parallellkopplade elektroniska formantresonatorer och insignalen var antingen ett sorl eller ett brus. Ungefär samtidigt introducerade Gunnar Fant den första seriella formantsyntesen, OVE I (Orator Verbis Electris), som bestod av seriekopplade formantresonatorer. Skillnaden mellan parallell och seriell formantsyntes beskrivs i stycket "Formantsyntes" nedan. Efter OVE I kom OVE II, OVE III och GLOVE, varav de två sistnämnda utvecklades på Kungliga Tekniska Högskolan och i vilka dagens högteknologiska talsyntessystem Infovox har sitt ursprung. Den första, elektroniska artikulatoriska syntesen, DAVO (Dynamic Analog of the VOcal tract), presenterades 1958 av George Rosen vid Massachusetts Institute of Technology (M.I.T.).

Det första fullständiga text-till-tal-systemet baserades på en artikulatorisk modell och utvecklades av Noriko Umeda m.fl. vid Electrotechnical Laboratory, Japan, 1968. Talet var monotont, men fullt begripligt. 1976 introducerade Kurzweil sin läsmaskin för blinda. Läsmaskinen innehöll en optisk scanner och läste relativt bra, men var alldeles för dyr för de flesta privatpersoner. Den användes på bibliotek och servicecentra för synskadade personer. I slutet av 1970-talet och början av 1980-talet introducerades en stor mängd text-till-tal-system. Bland dessa fanns MITalk utvecklat av Allen, Hunnicutt och Klatt samt Klattalk av Dennis Klatt. På 1980-talet började man använda datorer för att skapa talsyntes och idag sker i princip all talsyntesforskning med hjälp av datorer.

[1] Ordet spela används här eftersom man skötte pedaler med fötterna och spakar med händerna vilket kan liknas vid att spela orgel.

[2] Källa-filter-teorin säger om talproduktionen att uppkomsten av källsignalen vid ljudkällan och filtreringen av källsignalen i resonatorn är två av varandra oberoende led (Lindblad 1998).

1.3.2 Olika sorters talsyntes

Donovan (1996) delar in talsyntesmetoderna i två delar; systemmodellen, som försöker modellera den mänskliga talapparaten, och signalmodellen, som endast försöker modellera den resulterande talsignalen. Systemmodellen kallas ofta artikulatorisk syntes, medan signalmodellen delas in i formantsyntes och konkateneringssyntes.

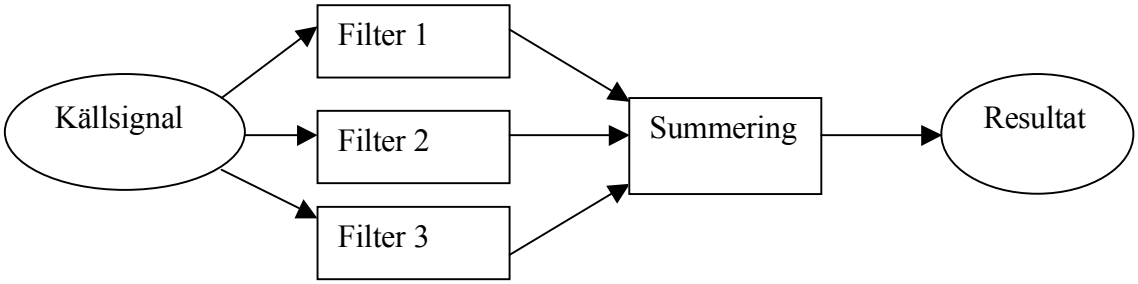
1.3.2.1 Artikulatorisk syntes

I artikulatorisk syntes används modeller av de mänskliga artikulatorerna (t.ex. tunga, läppar och käke) och stämbanden. Med hjälp av regler för vilka rörelsehinder de olika artikulatorerna har kan modellerna kontrolleras och flyttas mot målpositioner för varje fonem (Donovan 1996). Utveckling av artikulatorisk syntes verkar idag främst ske i syfte att forska om den mänskliga talapparaten och talproduktionen. Ett exempel är Haskins laboratories som har utvecklat ett artikulatoriskt syntesprogram, ASY, i syfte att studera förhållandet mellan talproduktion och talperception (Rubin & Goldstein 1995-8).

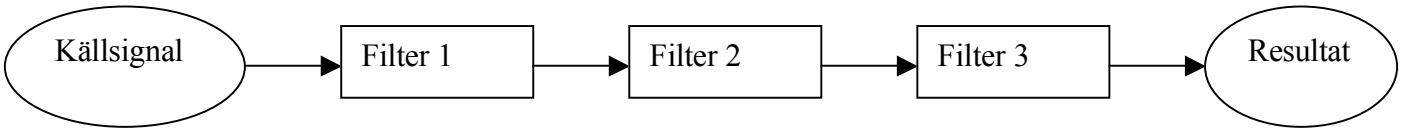
1.3.2.2 Formantsyntes

Formantsyntes är en sorts källa-filter-metod där ansatsröret modelleras utifrån ett antal resonanser som liknar formanterna i naturligt tal. För att producera begripligt tal behövs åtminstone 3 formanter och för att producera högkvalitativt tal används upp till 5 formanter. I ett formantfilter för varje formant specificeras både frekvens och bandbredd för aktuell formant. För att modellera ansatsröret kan man kombinera formantfiltren på två olika sätt; parallellt eller seriellt. Vid parallell formantsyntes appliceras källsignalen på alla formantfilter samtidigt och resultatet från varje formantfilter summeras sedan till ett slutresultat. Vid seriell formantsyntes appliceras källsignalen på ett formantfilter, vars resultat appliceras på nästa formantfilter osv. Resultatet från det sista formantfiltret utgör slutresultatet (Lemmetty 1999). Enligt Lass (1996) finns det klara fördelar med att använda parallell formantsyntes för klusiler och frikativor, medan seriell formantsyntes är bättre lämpat för vokaler och övriga sonoranter. Ofta används en kombination av de båda typerna, en sorts hybridsyntes som växlar mellan parallell och seriell formantsyntes, där den parallella slås på för klusiler och frikativor och den seriella för vokaler och övriga sonoranter. De två olika sorternas formantsyntes illustreras i Figur 2.

a)



b)



Figur 2. Parallell formantsyntes a) och seriell formantsyntes b).

1.3.2.3 Konkateneringssyntes

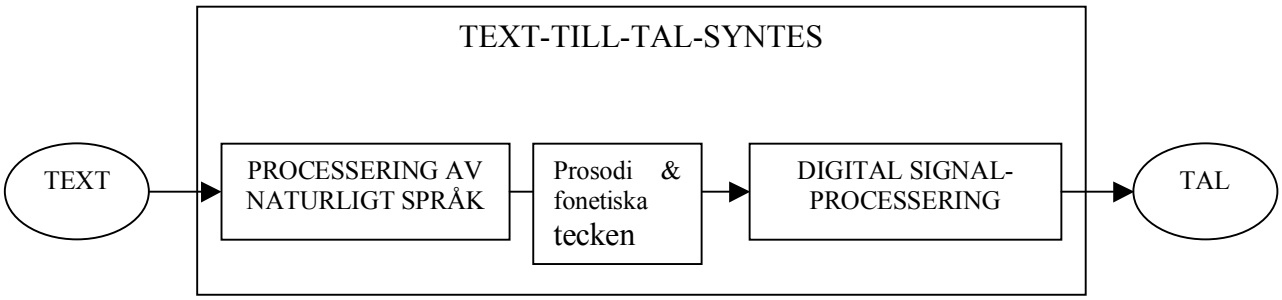
Vid konkateneringssyntes spelar man in talljudsenheter som sedan sätts ihop och bildar tal. Ljudenheternas längd kan variera mellan och inom olika konkateneringssynteser. Ord, stavelser, halvstavelser och difoner[3] är exempel på ljudenheter av olika längd som kan användas i konkateneringssynteser. Trots att synteser med långa enheter, som t.ex. ord, ofta låter mer naturliga än synteser med korta enheter, som t.ex. difoner, används oftast relativt korta ljudenheter vid konkateneringssynteser. Vid användning av ljudenheten ord måste alla ord som ska kunna syntetiseras spelas in och sparas, vilket kräver både mycket tid och mycket utrymme. Vid användning av difoner kan samma ljudenheter användas för att bygga upp olika ord, vilket resulterar i att ett mindre antal ljudenheter behövs (Donovan 1996). Difonen är idag en mycket vanlig ljudenhet inom konkateneringssyntes (Möhler 2001). Den kommer även att användas för att skapa en ny syntesröst inom denna uppsats.

1.3.3 Text-till-tal-syntes

En text-till-tal-syntes är ett datorbaserat system som ska kunna läsa upp vilken text som helst inom det språk som syntesen är skapad för. För detta behövs en komponent för processering av naturligt språk, NLP (Natural Language Processing), som producerar fonetiska tecken och

[3] Enhet som används vid konkateneringssyntes. Från två angränsande talljud mäts difonen från mitten av det första talljudet till mitten av det andra talljudet.

prosodi utifrån en given text, och en komponent för digital signalprocessering DSP (Digital Signal Processing), som omvandlar informationen från NLP till tal. I NLP finns bl.a. bokstav-till-ljud-regler, prosodigenerator samt analysatorer för kontext och morfologi. I DSP finns matematiska modeller och algoritmer (Dutoit 1997). I Figur 3 visas en bild av en generell text-till-tal-syntes.



Figur 3. Generell text-till-tal-syntes.

2. Ofelia _ en ny syntesröst

I detta avsnitt kommer jag att beskriva tillvägagångssätt, problem och resultat i mitt arbete med att skapa en ny syntesröst som jag valt att kalla Ofelia. Först beskrivs de hjälpprogram jag använt under arbetets gång; Festival, Festvox och MBROLA. Därefter kommer en beskrivning av hur det gick till när jag gjorde mitt pilotprojekt och till sist beskrivs arbetet med huvudprojektet.

2.1 Hjälpprogram

Här beskrivs de hjälpprogram som använts i arbetet med att skapa en ny syntesröst. De har varit avgörande för att kunna skapa syntesrösten inom ramen för denna uppsats.

2.1.1 Festvox

Festvox är ett projekt som syftar till att nya syntesröster ska skapas mer systematiskt och dokumenteras bättre. Ett mål är också att vem som helst (med lite kunskap i ämnet och rätt utrustning) ska kunna skapa en ny syntesröst. Festvox tillhandahåller ett paket innehållande dokumentation, redskap och program som underlättar arbetet med att skapa nya syntesröster. Paketet kan laddas ner från http://festvox.org/download.html och får brukas fritt för kommersiellt och icke-kommersiellt bruk. Festvox-paketet är utvecklat av Alan Black och Kevin Lenzo vid Carnegie Mellon University's speech group, Pittsburgh och är avsett att användas tillsammans med talsyntessystemet Festival (Black & Lenzo 1999-2001). Festvox-

paketet har i denna uppsats bl.a. använts vid framställning av difonlista samt vid inspelning och etikettering av difoner.

2.1.2 Festival

Festival är ett flerspråkigt talsyntessystem utvecklat av Alan Black, Richard Caley och Paul Taylor vid The Centre for Speech Technology Research, University of Edinburgh. Festival kan användas för att omvandla fonetiska tecken till talljud. Om det finns regler för omvandling från text till fonetiska tecken för den syntesröst man vill använda kan Festival användas för omvandling från text till tal. Man kan då skriva in text direkt till programmet eller anropa programmet med en textfil så läses texten upp. Festival kan laddas ner från http://www.cstr.ed.ac.uk/projects/festival/download.html och den senaste versionen får användas för både icke-kommersiellt och kommersiellt bruk (CSTR, University of Edinburgh 2001). Festival har använts för att lyssna på resultatet av pilotprojektet och senare också hela den nya syntesrösten.

2.1.3 MBROLA – MultiBand Resynthesis OverLap Add

MBROLA-projektet (Dutoit et al. 1996) startades av TCTL Lab, Faculté Polytechnique, Mons, Belgien. Dess mål är att främja akademisk talsyntesforskning och att anskaffa och sprida talsynteser för så många språk som möjligt för icke-kommersiellt bruk. Den som vill skapa en difondatabas kan spela in och segmentera difoner och skicka alla ljudfiler samt en indexfil till MBROLA-projektet som helt gratis processerar materialet och framställer en MBROLA-databas som skickas tillbaka. För detta krävs att man undertecknar ett avtal som säger att MBROLA-databasen får ingå i MBROLA-projektet för icke-kommersiellt och icke-militärt bruk. Avtalet säger vidare att den som tillhandahållit ljudfilerna och indexfilen får kommersiella rättigheter till databasen med villkoret att den används med MBROLA-projektets talgenerator MBROLA. MBROLA tar fonetiska tecken som insignal och omvandlar dessa till ljud. MBROLA-databasen kan även användas i talsyntessystemet Festival. MBROLA-projektet använder signalprocesseringsmetoden PSOLA (Pitch Synchronous OverLap and Add) som ofta ger en bättre kvalitet på syntesrösten än LPC (Linear Predictive Coding), som används i Festvox. (Dutoit 1996-2001)

2.2 Pilotprojekt

För att ta reda på om min egen röst var lämpad för inspelning av difoner till en syntesröst och för att upptäcka eventuella fällor och svårigheter i arbetet att skapa en difondatabas för svenska ville jag först testa med en liten del av språket. Ett sådant test skulle också visa om

inspelningsmiljön jag valt var bra. Min sydsvenska dialekt är en blandning av småländska och skånska eftersom jag växt upp och levt mina första 20 år utanför Tingsryd, söder om Växjö, i Småland och sedan flyttat till Lund där jag bott de senaste 4 åren. Eftersom det är min egen röst som ska användas vid inspelningen syftar "talaren" i fortsättningen på mig själv.

2.2.1 Val av talljud

För att fonetiska tecken skulle kunna användas på dator valdes istället för IPA (International Phonetic Alphabet) SAMPA-alfabetet (Speech Assessment Methods Phonetic Alphabet), en avbildning av IPA till en enklare teckenuppsättning som gör det lättare att använda på dator. MBROLA rekommenderar att man använder SAMPA-alfabetet som anger 46 talljud för det svenska språket (UCL Phonetics and Linguistics, University College London 1995-8). Av de 46 talljuden skulle en liten, representativ del väljas ut till pilotprojektet. Mängden talljud som valdes skulle representera så många av de olika sorters talljud som finns i svenskan som möjligt; långa och korta vokaler, glidljud, nasaler, frikativor och klusiler. Den skulle också vara tillräckligt stor för att skapa så många ord att man kunde avgöra om kvaliteten på pilotrösten var god. Mängden talljud fick dock inte bli för stor, eftersom antalet difoner, och därmed arbetet, ökar drastiskt för varje talljud som läggs till. Förutom att varje talljud i mängden bildar en difon med alla talljud i mängden förekommer också difonerna ’tystnad-talljud’ (ordinitialt) och ’talljud-tystnad’ (ordfinalt) för alla talljud i mängden. Slutligen valdes följande talljud till pilotprojektet:

Typ av talljud	IPA	SAMPA
lång vokal	ɪː	i:
lång vokal	ɑː	A:
kort vokal	ɪ	I
kort vokal	ʌ	a
nasal	m	m
frikativa	s	s
frikativa	ʃ	S[4]
frikativa	ç	C
glidljud	ʁ	r[4]
klusil	p	p

Tabell 1. IPA- och SAMPA-alfabetens beteckningar för pilotprojektets talljud.

[4] I den lista över symboler för svenska talljud i SAMPA-alfabetet som har använts (UCL Phonetics and Linguistics, University College London 1995-8) anges inga allofoner för dessa talljud. Därför har de som finns i listan och som även används för uppsvenska dialekter använts här, trots att talljuden i de sydsvenska dialekterna kraftigt skiljer sig från dem i de uppsvenska.

Detta ledde till 100 (10*10) vanliga difoner, 10 difoner av typen ’tystnad-talljud’ samt 10 difoner av typen ’talljud-tystnad’, dvs. totalt 120 difoner. I en speciell fil från Festvox specificerades de talljud som valts ut samt deras fonologiska särdrag. I en annan fil angavs vilka talljud som är vokaler, vilka som är konsonanter, vilka som förekommer i konsonantkluster, vilka som endast förekommer i codan[5] och vilka som endast förekommer i onset[6]. Med ett program från Festvox som använder dessa båda filer skapades sedan en lista av nonsensord innehållande de aktuella difonerna i en ny fil. Nonsensord, som inte har någon känd prosodi, används istället för riktiga ord för att talaren lättare ska kunna hålla en monoton ton och låta bli att betona några stavelser. Några av orden innehöll s.k. spegeldifoner, dvs. två difoner av typen 'talljud1-talljud2' och 'talljud2-talljud1'[7]. Filen innehöll därför 101 ord och inte 120.

2.2.2 Inspelning

Inspelningen gjordes direkt till hårddisk med hjälp av ett headset (modell Labtech Axis 302) som kopplades till en dator. Ett headset är bättre än en vanlig mikrofon eftersom avståndet mellan munnen och mikrofonen blir konstant, vilket reducerar risken för stora skillnader i ljudstyrka vid inspelningen av olika difoner. För att undvika det brusljud som uppstår från datorns ljudkort kunde inspelningen istället ha gjorts på DAT-band[8]. Det främsta skälet till att hårddisk valdes framför DAT-band var att inspelningen av nonsensord nu kunde ske till olika filer som därefter kunde etiketteras automatiskt. På så sätt slipper man alltså stycka upp talsignalen i olika ord, eftersom detta görs redan vid inspelningen. En annan fördel är att man direkt kan korrigera fellästa ord genom att läsa in dem igen och skriva över det fellästa ordet.

I Festvox finns program för att skapa s.k. suffleringsord (eng. prompts). Suffleringsord är ord syntetiserade med en annan syntesröst som används i dubbelt syfte. Man kan spela upp suffleringsordet för varje ord precis innan man spelar in samma ord för att hålla en jämn och monoton ton genom hela inspelningen. Detta är viktigt för att kvaliteten på syntesrösten ska bli bra. Man kan också använda suffleringsorden för en automatisk etikettering av de inspelade nonsensorden, vilket sparar mycket tid vid segmenteringen.

Vid inspelningen användes ett program från Festvox som spelar upp suffleringsordet för aktuellt ord och sedan direkt spelar in talaren som upprepar samma ord. Därefter spelas både

[5] Konsonant(er) som avslutar en stavelse.

[6] Inledande konsonant(er) i en stavelse.

[7] Nonsensordet "tatata" innehåller t.ex. spegeldifonerna 't-a' och 'a-t'.

[8] Digital Audio Tape. En digital lagringsenhet som ser ut ungefär som ett kassettband.

suffleringsordet och det nyinspelade nonsensordet upp så att man kan höra resultatet. Samtidigt som suffleringsordet spelas upp visas på skärmen vilket ord och vilken difon det är som ska spelas in. Om man inte är nöjd med det ord man just har spelat in kan man direkt avbryta programmet och starta från samma ord igen. Om man inte avbryter spelar programmet upp suffleringsordet för nästa ord och så fortsätter det. Programmet spelar in under två sekunder för varje ord, förutom vid riktigt korta ord där det spelar in under en sekund, så det gäller att vara med. Om man tycker att det går för fort kan man ändra hastigheten i inställningarna till programmet. Vid inspelningen kördes programmet med standardhastigheten.

Själva inspelningen tog ca 20 minuter och det var inte särskilt svårt att hålla en jämn och monoton ton under denna korta tid. Det svåraste var att snabbt bedöma om de nyinspelade orden var bra eller om något var fel eller otydligt så att inspelningen skulle avbrytas och startas från samma ord igen.

Rummet som användes för inspelning är ljudisolerat och det enda ljud som förekom var surret från datorn som användes vid inspelningen. För att uppnå ett minimalt bakgrundsljud sattes en ljudisolerande skiva upp mellan talaren och datorn. Skärmen och tangentbordet var på samma sida skivan som talaren så att denne kunde se vad som hände på skärmen och avbryta programmet när som helst genom att använda tangentbordet.

2.2.3 Segmentering

Vid segmenteringen spelar de redan nämnda suffleringsorden en stor roll. De har en korrekt etikettering eftersom de har syntetiserats utifrån talljudsetiketter. På bas av en akustisk jämförelse mellan de nyinspelade orden och de syntetiserade suffleringsorden sätts etiketter för de talljud som ingår in i de nyinspelade orden. Detta kallas inrättning (eng. aligning). Efter denna automatiska etikettering finns alla etiketter i rätt ord och i många fall på rätt plats. Det är sedan viktigt att gå igenom alla orden och manuellt flytta de etiketter som hamnat fel till rätt plats. För detta användes programmet *fvlab,* utvecklat av Johan Frid vid Institutionen för Lingvistik, Lunds universitet. Etiketterna sätts vid högra gränsen av sitt talljud. Vid klusiler kan man sätta in en extra etikett, '.cl', för att markera var själva explosionen börjar. Detta gör man för att undvika att ocklusionen kommer med i difonen, där endast halva talljudet används. Man kan också sätta in en extra difongräns, 'DB', om man vid något talljud inte vill att difonen ska mätas från mitten av talljudet. Detta kan vara aktuellt t.ex. vid difoner av typen ’vokal-tystnad', där energin i vokalen ofta avtar drastiskt. Om difonen som vanligt tas från mitten av vokalen, där energin redan hunnit sjunka rejält, kan en skarp skarv uppstå när difonen sätts ihop med en

annan difon som slutar på samma vokal. Detta beror på att vokalen i den andra difonen förmodligen har högre energi eftersom den tagits i början eller mitten av ett ord där energin normalt är större än i slutet.

2.2.4 Vågformsgenerering på två sätt

Med hjälp av Festvox kan man själv skapa en ljudsyntes. Med några enkla kommandon kan glottispulsmarkeringar extraheras och flyttas till närmaste topp i vågformen. Detta är i princip samma sak som att göra en grundtonsanalys. Skillnaden är att man istället för en exakt frekvens får fram var perioderna ligger och därmed dess längd. Här är programmet inställt för en manlig röst så för att få det att fungera för en kvinnlig röst krävs en ändring i inställningarna så att rätt frekvensområde undersöks. Man kan också mäta energin i vokalerna och få fram ett medelvärde som sedan används för att modifiera energin hos de vokaler som ligger långt från medelvärdet. Med hjälp av glottispulsmarkeringarna görs sedan en glottispulssynkron LPC-analys (Linear Predictive Coding) som är den metod för signalprocessering som Festvox använder. När allt detta var gjort kunde pilotrösten utvärderas. Genom att skriva in en rad fonetiska tecken som motsvarade ett ord eller en mening kunde man lyssna på syntesrösten som läste upp resultatet. Syntesrösten var förstås helt monoton och första intrycket var att den lät väldigt hes. Som ovan lyssnare till syntestal utan prosodi var det svårt att bedöma om röstkvaliteten var tillräckligt god för att jag skulle gå vidare med projektet. För att kvaliteten skulle kunna utvärderas ytterligare skulle difondatabasen skickas till MBROLA-projektet i Belgien för processering så att resultatet kunde jämföras med det som genererats med Festvox. MBROLA behövde samtliga ljudfiler samt en indexfil som såg lite annorlunda ut än den som automatiskt genereras med Festvox program. I den automatiskt genererade filen fanns en rad för varje difon med följande information:

difonnamn ljudfilsnamn tid1 tid2 tid3

Difonnamn visar vilken difon som avses (t.ex. s-r) och ljudfilsnamn är namnet på den ljudfil som innehåller aktuell difon. Tiderna visar var i ljudfilen difonens början (tid1), mitt (tid2) och slut (tid3) finns. Indexfilen som MBROLA ville ha skulle istället se ut på följande sätt:

ljudfilsnamn difonnamn tid1 tid3 tid2

Här anges tiderna i samplingspunkter (eng. sample points) istället för, som i den ursprungliga filen, sekunder. Samplingspunkter får man genom att multiplicera tiden i sekunder med

samplingsfrekvensen, som i detta fall är 16000. Här går det alltså 16000 samplingspunkter på en sekund. Förutom dessa ändringar skulle tystnadssymbolen '#' ändras till '_', '_' tas bort ur ljudfilsnamnen och filtypen läggas till efter filnamnen. När dessa ändringar gjorts skickades ljudfilerna och indexfilen till MBROLA som, efter att ha gjort några ändringar som jag hade missat, kunde processera materialet och skicka tillbaka en MBROLA-databas.

2.2.5 Resultat

Efter att ha lyssnat på den nya syntesrösten var jag lättad, eftersom det var en dramatisk skillnad jämfört med den syntesröst som genererats med Festvox. Den nya syntesrösten hade en klar och fin ton, vilket visade att det var väl värt att skicka ljudfilerna till MBROLA-projektet för processering. Jag bestämde mig för att gå vidare med en stor databas för hela språket utan att ytterligare experimentera med pilotrösten. En otydlighet i skillnaden mellan långa och korta vokaler iakttogs och jag hade detta i åtanke inför inspelningen av den stora databasen. Inga större problem påträffades under pilotprojektet, vilket tydde på att inspelningsmiljön, talarens röst och de metoder som använts fungerade bra och kunde användas i huvudprojektet.

2.3 Huvudprojekt

2.3.1 Val av talljud

När man skapar en ny syntesröst är det viktigt att fundera över vad man vill att den ska kunna användas till. En svensk syntesröst bör kunna säga alla svenska ord, men hur är det med t.ex. utländska namn? I svenska texter förekommer ofta utländska (särskilt engelska) namn på orter och människor. Om man vill att syntesrösten ska kunna uttala sådana namn bör man därför ta med vissa talljud som inte förekommer i svenskan, men som är vanliga i t.ex. engelska, s.k. xenofoner. För en diskussion om svenska xenofoner se Eklund & Lindström (2001). I denna uppsats har jag begränsat syntesrösten till svenska språket. Detta eftersom jag inte ansåg behovet av en syntesröst som kan uttala utländska namn tillräckligt stort jämfört med det arbete som skulle tillkomma med dessa utländska talljud. I SAMPA-alfabetet anges 46 talljud för svenska språket. Däribland finns supradentalerna[9] 'rt', 'rd', 'rn', 'rs' och 'rl' som inte förekommer i talarens sydsvenska dialekt. Dessa togs därför inte med i listan över talljud för detta projekt. I SAMPA-alfabetet görs även en distinktion mellan ä-ljuden i orden "rätt" och "vett". Denna distinktion finns inte hos talaren och ljuden räknas därför som ett och betecknas med 'e' i detta projekt. Något som saknas i SAMPA-alfabetets lista över svenska talljud är de aspirerade,

[9] När r följs av t, d, n, s eller l assimileras r i de flesta uppsvenska dialekter med efterföljande konsonant och en supradental eller retroflex bildas, t.ex. bord, fors.

tonlösa klusilerna som är allofoner av de icke-aspirerade och förekommer stavelseinitialt. Dessa finns med i detta projekt och betecknas med 'ph', 'th' och 'kh'. Antal talljud i projektet blev till slut 43. Detta ledde till 1849 (43*43) vanliga difoner, 43 difoner av typen 'tystnad-talljud' och 43 difoner av typen 'talljud-tystnad', d.v.s. totalt 1935 difoner. Eftersom vissa talljud inte förekommer i alla kontexter kunde ett antal difoner uteslutas. Ng-ljudet förekommer t.ex. endast i codan och de aspirerade, tonlösa klusilerna samt h förekommer endast i onset. Alla konsonanter ingår heller inte i konsonantkluster. Allt detta specificerades i en särskild fil som användes tillsammans med en fil innehållande en lista över alla talljuden och dess egenskaper för att skapa en ny fil med en lista av nonsensord innehållande alla difonerna. Listan innehöll 1439 nonsensord med totalt 1658 difoner.

Några av tecknen som används i SAMPA-alfabetet fungerar inte att använda i Festival och ytterligare några fungerar inte att använda i segmenteringsprogrammet fvlab. Dessa tecken byttes därför ut så att Festival och fvlab skulle kunna användas utan problem.

I Bilaga 1 finns en lista över projektets talljud och dess SAMPA-tecken. Där anges också vilka SAMPA-tecken som bytts ut och mot vad.

2.3.2 Inspelning

Inspelningen skedde på samma sätt, med samma hjälpmedel och i samma miljö som vid pilotprojektet. Tiden för inspelningen var totalt ca 6 timmar. Under denna tid tog talaren flera pauser för att dricka, äta och röra på sig. Det var betydligt svårare att hålla en jämn och monoton ton än under pilotprojektet, eftersom många fler ord skulle spelas in. Suffleringsorden var till stor hjälp, eftersom de har en monoton och jämn prosodi som talaren kunde imitera.

2.3.3 Segmentering

Segmenteringen gjordes, precis som vid pilotprojektet, med programmet fvlab och tog ca 25 timmar. Under segmenteringen gjordes också en ordentlig genomlyssning av de inspelade nonsensorden, varvid vissa ord med mindre bra kvalitet upptäcktes. Den dåliga kvaliteten bestod främst i knarr i vokalljud, långa vokaler som blivit korta och tvärtom samt fall då endast en konsonant uttalats vid konsonantrepetition. Detta resulterade i att 60 ord spelades in och segmenterades på nytt.

2.3.4 Vågformsgenerering

Med Festvox-paketet gjordes en LPC-syntes som fick ungefär samma ljudkvalitet som pilotrösten. Sedan skulle även dessa ljudfiler skickas till MBROLA-projektet för processering. P.g.a. det stora antalet ord skapades ett Java-script[10] som utförde de flesta av ändringarna i indexfilen. Denna gång ändrades också ljudfilernas format från wav till raw[11] eftersom det framkommit att det wav-format som används av MBROLA-projektet är ett annat än det som hittills hade använts i detta projekt. Filerna skickades till MBROLA-projektet och en MBROLA-databas erhölls. Röstkvaliteten var likvärdig med den i pilotprojektet, men hade en något lägre volym, vilket inte orsakar några problem och vars orsak är okänd.

2.3.5 Från text till tal

För att till Festival direkt kunna skriva in text för omvandling till tal krävs en definition av vilka ljud bokstäverna i texten motsvarar. Detta kan göras genom att definiera ett lexikon där alla ord (inklusive transkription) som ska kunna syntetiseras räknas upp. Ett annat sätt är att definiera regler för vilka ljud de enskilda bokstäverna motsvarar i olika kontexter. Jag valde att definiera ett antal bokstav-till-ljud-regler som kompletterades med ord som utgjorde undantag från reglerna. I Festvox-paketet finns fördefinierade filer för just detta, där man direkt kan skriva in regler och lexikon i ett förbestämt format. Genom tillgång till motsvarande filer för en syntesröst som tidigare skapats på institutionen av Johan Frid underlättades arbetet. Johan Frids regler var skrivna för en skånsk syntesröst och hans talljudsmängd var inte identisk med min. Dessa regler användes som grund för den nya syntesrösten, med några ändringar som eliminerade de skånska dragen och några tillägg p.g.a. fler talljud. Ett utdrag ur filen med reglerna finns i Bilaga 2.

2.3.6 Resultat

Resultatet av arbetet blev en ny syntesröst med en bra röstkvalitet och en sydsvensk dialekt. Syntesrösten klarar alla talljud som finns i talarens dialekt och kan därmed uttala alla svenska ord. Med hjälp av bokstav-till-ljud-reglerna kan text skrivas in direkt till Festival som läser upp resultatet. Bokstav-till-ljud-reglerna kan förbättras. I nuvarande skick klarar syntesen inte av att uttala alla ord på ett korrekt sätt. Prosodin är i det närmaste obefintlig och här finns mycket att göra för den som har tid och lust. När några av institutionens fonetiker lyssnade på syntetiserade

[10] Textfil som innehåller en rad instruktioner skrivna i programmeringsspråket Java. Instruktionerna utförs när Java-scriptet exekveras.

[11] MBROLA-projektet vill ha 16 bit little endian, vilket är ett sätt att strukturera ljudfiler.

meningar utan prosodi tyckte de att det var svårt att uppfatta vad som sades. Detta berodde främst på att reglerna är enkla och många ord blir felbetonade. När en mening med prosodi demonstrerades var intrycket betydligt bättre. Några tydliga skarvar mellan difonerna upptäcktes inte.

3. Sammanfattning

Pilotprojektet fyllde sitt syfte genom att visa att talarens röst, inspelningsmiljön och de metoder jag valt att använda fungerade bra. Det visade också det stora värdet av att skicka ljudfilerna till MBROLA och låta dem producera en difondatabas, eftersom röstkvaliteten från denna databas blev betydligt bättre än från den databas som producerats med Festvox. Några otydligheter i inspelningarna upptäcktes så att talaren kunde uppmärksammas på dessa. Allt detta tillsammans med det faktum att hela arbetet redan gjorts en gång i liten skala underlättade betydligt arbetet med att skapa en syntesröst för hela språket. I huvudprojektet var arbetet med att välja vilka talljud som skulle ingå i difondatabasen och att välja lämpliga beteckningar för dem som inte var självklara större än i pilotprojektet. Här valdes utländska och uppsvenska talljud bort, medan några andra talljud som ingår i talarens sydsvenska dialekt lades till. Genom skapandet av några enkla bokstav-till-ljud-regler blev det möjligt att direkt skriva in text till Festival istället för, som vid pilotprojektet, fonetiska tecken. Beträffande prosodin har mycket lite gjorts och mycket kan göras för att förbättra den. Hjälpprogrammen har varit till mycket stort stöd i arbetet. Utan Festvox och Festival hade arbetet försvårats och framför allt tagit mycket längre tid. Utan MBROLA hade en databas kunnat framställas på samma tid men med ett betydligt sämre resultat i röstkvaliteten. På MBROLA-projektets webbsida kan man snart lyssna på en provmening av den första kvinnliga syntesrösten med sydsvensk dialekt.

Referensförteckning

Black, Alan W. & Lenzo, Kevin A. (1999-2001). *Festvox*. Hämtat från http://festvox.org/ den 3/12 2001.

CSTR, University of Edinburgh (2001). *The Festival Speech Synthesis System*. Hämtat från http://www.cstr.ed.ac.uk/projects/festival/ den 3/12 2001.

Donovan, Robert Edward (1996). *Trainable speech synthesis*. Cambridge University Engineering Department.

Dutoit, Thierry (1997). *An introduction to text-to-speech-synthesis* (Text, Speech and Language Technology). Kluwer Academic Publishers, Dordrecht.

Dutoit, Thierry (1996-2001). *The MBROLA Project*. Hämtat från http://tcts.fpms.ac.be/synthesis/mbrola.html den 3/12 2001.

Dutoit et al. (1996). *The MBROLA projekt: Towards a set of high quality speech synthesizers free of use for non-commercial purposes. Proceedings ICSLP '96 (*Philadelphia) 3, 1393-96.

Eklund, Robert & Lindström, Anders (2001). *Xenophones: An investigation of phone set expansion in swedish and implications for speech recognition and speech synthesis.*

Lemmetty, Sami (1999). *Review of Speech Synthesis Technology*. Hämtat från http://www.acoustics.hut.fi/~slemmett/dippa/contents.html den 3/12 2001.

Möhler, Gregor (2001). *Examples of Synthesized Speech*. Hämtat från http://www.ims.uni-stuttgart.de/~moehler/synthspeech/examples.html#swedish den 3/12 2001.

Rubin, Philip & Goldstein, Louise (1995-8). *Articulatory synthesis*. Hämtat från http://www.haskins.yale.edu/Haskins/MISC/ASY/asy.html den 3/12 2001.

Traunmüller, Hartmut (1997-2000). *Wolfgang von Kempelen's and the subsequent speaking machines*. Hämtat från http://www.ling.su.se/staff/hartmut/kemplne.htm den 3/12 2001.

UCL Phonetics and Linguistics, University College London (1995-8). *SAMPA for Swedish*. Hämtat från http://www.phon.ucl.ac.uk/home/sampa/swedish.htm den 3/12 2001.

Bilaga 1

SAMPA-tecken	Alternativt tecken	Exempelord	Transkription
p		spik	spi:k
	ph	pil	phi:l
b		bil	bi:l
t		stal	tA:l
	th	tal	thA:l
d		dal	dA:l
k		skal	skA:l
	kh	kal	khA:l
g		gås	go:s
f		fil	fi:l
v		vår	vo:r
s		sil	si:l
S		sjal	SA:l
C		tjock	COk
h		hal	hA:l
m		mil	mi:l
n		nål	no:l
N		ring	rIN
r		ris	ri:s
l		lag	lA:g
j		jag	jA:g
i:		vit	vi:t
e:		vet	ve:t
E:		säl	sE:l
y:		syl	sy:l
}:	>:	hus	h>:s
2:	oe:	föl	foe:l
u:		sol	su:l
o:		hål	ho:l
A:		hal	hA:l
I		vitt	vIt
e		vett	vet
Y		bytt	bYt
u0	>	buss	b>s
2	oe	föll	foel
U		bott	bUt
O		håll	hOl
a		hall	hal
{:	<:	här	h<:r
9:	oer:	hör	hoer:r
{	<	herr	h<r
9	oer	förr	foerr
@		pojken	phOjk@n

Bilaga 2

Ord som är undantag från reglerna:

```
(lex.add.entry '("keps" nn (((k e p s) 1))))
(lex.add.entry '("choklad" nn (((S U) 0) ((k l A: d) 1))))
(lex.add.entry '("chock" nn (((S O k) 1))))
(lex.add.entry '("chiffer" nn (((S I f) 1) ((@ r) 0))))
(lex.add.entry '("och" kn (((O k) 0))))
(lex.add.entry '("människor" nn (((m e) 1) ((n I) 0) ((S U r) 1))))
```

Symboler:

```
(lex.add.entry '("@" n (((s n A:) 1) ((b e l) 0) ((A:) 1))))
(lex.add.entry '("%" n (((p r U) 0) ((s e n t) 1))))
(lex.add.entry '("+" n (((p l > s) 1))))
```

Definition av hårda och mjuka vokaler samt konsonanter:

```
;; vowels
(V a e i o u y å ä ö)
;; hard vowels (hårda)
(Vh a o u å)
;; soft vowels (mjuka)
(Vm e i y ä ö)
;; consonants
(C b c d f g h j k l m n p q r s t v w x z)
```

Bokstav-till-ljud-regler:

```
( [ a ] r C = A: )  ; karta, karlar
( [ a ] r = a )  ; karamell, stolar
( [ a ] C C = a )  ; katter
( [ a ] C # = A:1 )  ; tal
( [ e ] l # = @ )  ; cykel
( [ e ] n # = @ )  ; pojken
( [ e ] r # = @ )  ; böcker
( [ e ] C C = e )  ; vett
( [ e ] C # = e:1 )  ; vet
( [ i ] C C = I )  ; vitt
( [ i ] C # = i:1 )  ; vit
( [ o ] C C = U )  ; bott
( [ o ] C # = u:1 )  ; bor
( [ u ] C C = > )  ; buss
( [ u ] C # = >:1 )  ; bus
( [ y ] C C = Y )  ; nyss
( [ y ] C # = y:1 )  ; nys
( [ å ] C C = O )  ; håll
( [ å ] C # = o:1 )  ; hål
( [ ä ] r C = < )  ; skärm
( [ ä ] C C = e )  ; herr
( [ ä ] r # = <:1 )  ; här
( [ ä ] C # = E:1 )  ; häl
( [ ö ] C C = oe )  ; föll
( [ ö ] r C = oer )  ; förr
( [ ö ] C # = oe:1 )  ; föl
```

```
( [ ö ] r # = oer:1 )  ; för
( [ a g e ] # = A:1 S )  ; garage, plantage

;;; sj-sound
( [ s i ] o n = S )  ; fusion
( [ t i ] o n = S )  ; station
( [ s t j ] = S )  ; stjärna
( [ s k j ] = S )  ; skjorta
( [ s c h ] = S )  ; schema
( [ s h ] = S )  ; shorts
( [ s k ] Vm = S )  ; skära
( [ s j ] = S )  ; sjal
;;; ng-sound
( [ n k ] = N k )  ; bänk
( [ n g ] = N )  ; klang
( [ g n ] = N n )  ; ugn
;;; tj-sound
( [ t j ] = C )  ; tjata
( [ k j ] = C )  ; kjol
( [ c h ] = C )  ; chips, chok OBS! undantag: choklad, chiffer mm
;;; g is j sometimes
( l [ g ] e = j )  ; älgen
( l [ g ] # = j )  ; älg
( r [ g ] e = j )  ; färgen
( r [ g ] # = j )  ; färg
( [ c ] Vh = k )  ; café
( [ c ] Vm = s )  ; centrum
( [ c k ] = k )  ; tacka
( [ c c ] = k s )  ; accelerera
( # [ d j ] = j )  ; djur
( # [ h j ] = j )  ; hjul
( # [ l j ] = j )  ; ljud
( # [ g j ] = j )  ; gjort
( # [ k ] Vm = C )  ; köra OBS! undantag: keps
( # [ g ] Vm = j )  ; get
```