Akustiske databaser for norsk

Innhold

Om desse databasane

Dei akustiske databasane skildra under vart utvikla av firmaet Nordisk språkteknologi holding AS (NST), som gjekk konkurs i 2003. I 2006 kjøpte eit sameige samansett av Universitetet i Oslo, Universitetet i Bergen, Noregs teknisk-naturvitskaplege universitet, Språkrådet og IBM konkursbuet etter NST, for å syte for at dei språklege ressursane som NST hadde utvikla, vart ivaretekne. Nasjonalbiblioteket fekk i oppdrag frå Kulturdepartementet å byggje opp ein norsk språkbank i 2009, og starta dette arbeidet i 2010.

Ressursane etter NST vart overførde til Nasjonalbiblioteket i mai 2011, og dei vert no gjort tilgjengelege i Språkbanken, førebels utan vidare handsaming. Språkbanken er open for attendemeldingar og forslag frå brukarane til korleis ressursane kan verte forebtra, og tek òg gjerne imot forbetra versjonar av databasane som brukarar ønskjer å dele med andre brukarar gjennom Språkbanken. Respons og attendemeldingar kan sendast til sprakbanken@nb.no.

Teksta i skildringa som følgjer er i sin heilskap skrive av Gisle Andersen, og henta frå rapporten *Gjennomgang og evaluering av språkressurser fra NSTs konkursbo*, skrive i 2005 før sameiget (sjå over) kjøpte konkursbuet etter NST, og er ein fagleg og teknisk gjennomgang av ressursane. Teksta er tillempa situasjonen slik han er i dag. Gisle Andersen har gitt Språkbanken løyve til å nytte teksta. Den nemnde rapporten kan lastast ned i sin heilskap saman med annan informasjon via lenkja http://www.nb.no/sbfil/dok/dok.tar.gz.

Nasjonalbiblioteket, juni 2011

1. Akustiske databaser for talegjenkjenning

De akustiske databasene omfatter to hovedtyper: databaser produsert for talegjenkjenning/ diktering, og databaser produsert for talesyntese. Førstnevnte kategori er den desidert mest omfattende.

I avsnitt 2 følger en generell beskrivelse av NSTs databaser for talegjenkjenning/diktering. Deretter følger en språkspesifikk ressursoversikt for norsk/svensk/dansk i avsnitt 3, med beskrivelser av ressursenes omfang, formater og valideringsgrad. Avsnitt 4 gir en beskrivelse av den metode og standard som ligger til grunn for validering av opptakene. Avsnitt 5 og 6 tar for seg spesifikke språklige tema og dialektområder. Deretter følger en samlet kvalitativ vurdering av de akustiske ressursene i avsnitt 7. Akustiske databaser for talesyntese beskrives i avsnitt 8.

2. Generelt om NSTs akustiske databaser

Materialet fordeler seg på ulike kategorier definert av formålet for innspillingen. Disse er beskrevet nedenfor.

Generelt for databasen gjelder at den er i sin helhet samlet inn og validert av NST selv. Unntak fra dette er deldatabasene SpeechDat og Telia, som er innkjøpte baser og del av såkalte ”in kind”-ressurser som NST anskaffet ved aksjeemisjon. SpeechDat består av data for mobiltelefoni og fasttelefoni for norsk, svensk og dansk. NST har ikke eierrettighetene til dette materialet, og det er således ikke omtalt i påfølgende avsnitt. Denne ressursen er godt dokumentert andre steder (jf. http://www.speechdat.org/) Telia-materialet består av kontormiljøbaserte innspillinger for talegjenkjenning gjort i Stockholmsområdet.

Hoveddelen av de akustiske ressursene er spilt inn og validert ved hjelp av programvare fra L&H. Opptaksprogrammet DSDR (Desktop Speech Digital Recorder) er benyttet, så sant ikke noe annet er opplyst nedenfor. All validering har foregått ved hjelp av valideringsprogrammet DSVS (Desktop Speech Validation Station).

Tilgang til denne proprietære programvaren er imidlertid ingen forutsetning for å nyttiggjøre seg dataene til fremtidige formål. Selve dataene foreligger nemlig i generelt anvendelige formater, i form av lydfiler i PCM/wav-format og spl-loggfiler i rent tekstformat. (Mer utførlig beskrivelse følger.) Lydfilene ligger lagret i ukomprimert form (ikke bruk av zip eller tilsvarende verktøy).

I det følgende gis en kvantitativ og kvalitativ beskrivelse av databasens innhold.

3. Akustiske databaser for norsk

3.1. Opptak for talegjenkjenning (ASR/diktering), 16 kHz

Filene i denne deldatabasen kan lastes ned via lenkene under. På grunn av det store volumet, er materialet delt opp:

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-1.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-2.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-3.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-4.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0464-testing.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dyf.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dym.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.validering.tar.gz

Deldatabasen ADB_OD_Nor.NOR er samlet inn for produksjon av teknologi for akustisk modellering for PC/Multimedia talegjenkjenning og for automatisk diktering (Office ASR and Dictation). Opptakene er gjort i lukket kontormiljø og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av setninger fra NSTs norske korpus. Databasen består av en treningsdel og en testdel, der førstnevnte brukes til å trene selve den akustiske modellen, mens sistnevnte brukes for testformål. Fordelingen av opptak for de to delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Str (GB)
Trening	nor0463	312	900	280800	97,5
Testing	nor0464	987	80	78960	26,9

Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innpilt enhet, dvs. som oftest en setning eller noen tilfeller en frase eller enkeltord. Databasen er organisert etter en bestemt katalogstruktur, som vist nedenfor.

Lydfiler:	D:\adb_0464\speech\scr0464\23\04642301\r4640007
Annoteringsfil:	D:\adb_0464\data\scr0464\23\04642301\r4640007.spl
Liste over annoteringsfiler:	D:\adb_0464\doc\Spl.lst
Instruksjoner til informant:	D:\adb_0464\doc\nor464.scr

Konvensjonene for navngiving av spl-filene i katalogen data er som følger: …/skriptnummer/stasjonsnummer/gruppenummer/loggfil. Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding:	lineær PCM
Filformat:	ukodete rådata (headerless raw)
Samplingsfrekvens:	16 kHz
Oppløsning:	16 bit
Format:	Intel PCM
Kanaler:	2 (stereo)

Dette formatet er i overensstemmelse med kravspesifikasjonene fra L&H. En del av materialet er konvertert fra dette utgangspunktet til et format som kreves for produksjon av akustiske modeller basert på IBM-teknologi. Disse opptakene ligger i følgende arkiver:

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dyf.tar.gz (kvinner)
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dym.tar.gz (menn)

Denne deldatabasen består av de 312 treningsopptakene fra 415 menn og 485 kvinner. For dette materialet gjelder følgende teknisk informasjon:

Signalkoding:	lineær PCM
Filformat:	ukodete rådata (headerless raw)
Samplingsfrekvens:	16 kHz
Oppløsning:	16 bit
Format:	Motorola PCM
Kanaler:	1 (mono)

Innspillingsskriptet som treningsdataene bygger på har en dikteringsdel og en ASR-del. Førstnevnte del består av ordinære korpusekstraherte setninger som kreves til generelle dikkteringsformål, og utgjør manuskriptets 222 første enheter (setninger). Sistnevnte del omfatter de siste 90 enhetene og består av fraser som utgjør personnavn, stedsnavn, enkeltord, akronymer og annet som kreves til spesifikke talegjenkjenningsformål (ASR). Tegnsetting er eksplisitt lest.

Innspillingsskriptet som testdataene bygger på har en tilsvarende inndeling i en dikteringsdel og en ASR-del. Dikteringsdelen utgjør manuskriptets første 750 enheter, mens ASR-delen utgjør de siste 237 enhetene.

Hele materialet har blitt validert etter de kriterier og metoder som er nevnt i avsnitt 1.5. Valideringsfilene kan lastes ned via denne lenken:

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.validering.tar.gz

3.2. Opptak for diktering, 22 kHz

Denne deltatabasen kan lastes ned via følgende linker:

- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.mnt.data.nstdata.ambe.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.mnt.data.nstdata.norskdikterin g.ambe.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.skript-doc.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.tar.gz

Deldatabasen ADB_D_IBM-N er samlet inn for produksjon av teknologi for akustisk modellering for automatisk diktering (desktop). Ulikt foregående ressurs er opptakene spilt inn ved hjelp av IBM-programvaren ObjectRexx.3 Opptakene ble gjort i forbindelse med oppstart av samarbeidet mellom NST og IBM som ledd i opplæringsperioden av NST-ansatte. Databasen består

av tre deler innspilt til ulike formål: en testdel, en treningsdel og en modelleringsdel. Fordelingen av opptak i delene er som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Str (GB)
Modellering	mod	260	104	27040	6,24
Testing	test	160	20	3200	0,90
Enrollment	enroll	156	20	3120	0,90

Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding:	lineær PCM
Filformat:	ukodete rådata (headerless raw)
Samplingsfrekvens:	22 kHz
Oppløsning:	16 bit
Format:	Motorola PCM
Kanaler:	1 (mono)

Opptakene er gjort i lukket kontormiljø, og baserer seg på fonetisk balanserte manuskript, produsert på grunnlag av avistekst fra Aftenpostens 1996-årgang. Opptakene ligger som én lydfil per manuskriptlinje, som tilsvarer en innspilt enhet (setning, frase, enkeltord, tallrekke, bokstavrekke).

Denne deldatabasen er ikke validert. Det foreligger derfor begrenset dokumentasjon. Mikrofonen som ble benyttet er en Andrea NC-61, og lydkortet er Turtle Beach Montego II.

3.3. Opptak for telefoni, 8 kHz

Denne deldatabasen kan lastes ned via følgende lenker:

- http://www.nb.no/sbfil/talegjenkjenning/08kHz/nor.telefon.nsb.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/08kHz/nor.telefoni.original.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/08kHz/nor.telefoni.validering.tar.gz

Deldatabasen ADB_T_Nor.NOR inneholder opptak til telefoni, fordelt på fastlinje og mobiltelefoni. Disse er egnet til bruk ved produksjon av talegjenkjenningsteknologi for telefoni. Materialet er ikke inndelt i test- og treningsdata. NST har fulgt de generelle SpeechDat II-prosedyrene under opptak. Opptakene er gjort delvis med L&Hs programvare og delvis ved UMS Diginform.

Informantene har fått oppgitt telefonnummer og ringt inn og lest opp setninger. Opptakene inneholder 17 ytringer som er spontan tale i form av svar på spørsmål, og 40 ytringer med oppleste manuskriptsetninger. Skriptet nor0531.scr ble brukt både til fasttelefoni og mobiltelefoni. I forbindelse med utvikling av en ASR-applikasjon for NSB ble det også gjort opptak av navn på norske jernbanestasjoner. Disse er omfattet av manuskriptet nor0666.scr. Tabellen viser fordelingen av disse opptakene.

Formål	Skript	Linjer	Personer	Opptak	Str GB
Fasttelefoni	nor0531.scr	57	6231	355167	27,2
Mobiltelefoni	nor0531.scr	57	2018	115026	
Fasttelefoni	nor0666.scr	101	65	6565	0,4
Mobiltelefoni	nor0666.scr	101	37	3737	

Følgende teknisk informasjon gjelder for denne deldatabasen:

Signalkoding:	mu-Law
Filformat:	wav
Samplingsfrekvens:	8 kHz
Oppløsning:	16 bit
Format:	8-bit mu-Law Compressed
Kanaler:	1 (mono)

Deldatabasen knyttet til NSB-prosjektet finnes også her. I dette lagringsformatet foreligger dataene i IBM-kompatibelt telefoniformat, og følgende spesifikasjon gjelder:

Signalkoding:	A-Law
Filformat:	wav
Samplingsfrekvens:	8 kHz
Oppløsning:	16 bit
Format:	8-bit A-Law Compressed
Kanaler:	1 (mono)

Denne deldatabasen er kun delvis validert. Validering av 3108 fastlinje- og 1596 mobilopptak er foretatt, se lenken over.

Under valideringen har en del filer har blitt lagt til side fordi de er av for dårlig kvalitet; disse fins i underkatalogen …\forkastede opptak. En del opptak har blitt lagt til side av andre grunner enn dårlig kvalitet, f.eks. for å sikre optimal fordeling av informantene. Disse er ikke validerte og ligger i underkatalogen …\Opptak på vent. Ferdig validerte filer ligger i underkatalogen …\Validering. I tillegg finnes det cirka 1000 opptak som ikke har vært igjennom valideringsprosessen i det hele tatt.

Hele NSB-delen av databasen er validert.

3.4. Database med innspilte nølelyder

Denne deldatabasen kan lastes ned via følgende lenke:

- http://www.nb.no/sbfil/talegjenkjenning/nor.nolelyder.tar.gz

En deldatabase ble samlet inn for produksjon av spesifikke modeller for nølelyder, dvs. ikke-verbale lyder som uttales når en taler nøler mellom ord. Nølelydene omfatter en nasal og en vokal transkribert som <mmm> eller <eeeh> i manuskriptene. Et slikt materiale er et tillegg som brukes ved produksjon av generelle dikteringssystemer. Materialet ble spilt inn sammen med databasen ADB_OD_Nor.NOR beskrevet i avsnitt 1.2.1. De samme tekniske spesifikasjonene gjelder for dette subsettet. I likhet med resten av ADB_OD_Nor.NOR består det av en treningsdel og en testdel, som vist i tabellen.

Formål	Skript	Linjer	Personer	Opptak	Str (GB)
Trening	nor4631	28	10	200	0,6
Testing	nor4641	58	2	100	

Treningsmanuskriptet består av 20 vanlige setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd. Testmanuskriptet består av 50 vanlige setninger med to nølelyder i hver setning samt fire isolerte repetisjoner av hver nølelyd.

4. Validering

Begrepet ”validering” står sentralt i NSTs arbeid med akustiske ressurser. Felles valideringsprosedyrer er benyttet for alle innsamlingsprosjektene nevnt i avsnitt 1.2-1.4, og disse er i samsvar med prosedyrer for validering vedtatt av L&H. Valideringen er foretatt av språkassistenter som har arbeidet tett sammen i grupper, under faglig koordinering av gruppeledere.

Valideringen innebærer gjennomlytting av hvert enkelt opptak, merking av taledelens utstrekning, kontroll av samsvar mellom ortografi og uttale, merking av ikke-verbale hendelser og bakgrunnsstøy, merking av feiluttale og dialektformer, og angivelse av opptakets generelle tekniske og lingvistiske kvalitet.

Ikke-verbale hendelser (”events”) angis ved et finitt sett av koder eller markører, definert som følger:

SPK: Angir tydelige lyder som ikke er tale, laget av taleren, som hosting, kremting, pusting, spyttlyder, osv. Markeres ikke dersom de forekommer ord-medialt.

FRA: Brukes ved feil uttale, repetisjoner eller ved forekomst av nonsensord. Markøren [FRA] plasseres foran ordet som er feil uttalt.

INT: Brukes ved tidsavgrensete lyder, som musikk, andre stemmer, sirener, osv. [INT] plasseres der lyden forekommer, ev. foran ordet ved ordmedial forekomst.

STA: Brukes ved gjennomgående lyd, f. eks. sirener. Markøren [STA] plasseres der lyden høres for første gang.

TRC: Brukes ved avkortete (trunkerte) setninger, enten i begynnelsen eller slutten av signal som er blitt avkortet. Dette kan forekomme hvis informanten begynner å lese for tidlig, eller hvis en innringer legger på for tidlig.

FIL: Brukes ved fylte pauser, altså nølelyder. Markøren [FIL] plasseres hvor nølelyden forekommer i setningen.

DST: Brukes ved telefonforstyrrelser. Markøren [DST] plasseres foran ordet som blir forvrengt.

DIT: Brukes bare i tilfeller hvor en pipetone høres på begynnelsen av signalet. Denne er et signal til informanten om at han/hun kan begynne å snakke, og skal ikke være en del av opptaket.

På grunnlag av disse annoteringene kan man ved modellering basert på databasene sørge for at dataene som inngår i den akustiske modellen kun inneholder menneskelig tale og ikke forstyrrende signaler, og dette er med på å høyne den akustiske modellens kvalitet.

Kvalitetsmerkingen av dataene baserer seg på observasjoner av både tekniske og lingvistiske forhold. En skala fra A til E er benyttet, der A er angir beste kvalitet og E angir forkastete opptak. Kvalitetskriteriene er for øvrig som i tabellen på neste side.

Kontrollert innpust eller normal smatting regnes ikke som støy og er ikke markert som sådan. Det er heller ikke markert ved ubetydelig susing/knitring; litt lyd i bakgrunnen betraktes som normalt. Se avsnitt 1.7 for en mer detaljert beskrivelse av lingvistiske kvalitetskriterier.

A. Dersom man ikke hører andre ting enn talen gis karakteren A.

B. Dersom man observerer ikke-verbale hendelser i bakgrunnen (andre som snakker, spyttlyder, bakgrunnsstøy, osv.) gis karakteren B.

C. Dersom disse bakgrunnslydene er veldig tydelige, men ikke overdøver talen, gis karakteren C. C gis også dersom det er mindre enn 100 ms mellom begynnelsen eller slutten av talesignalet og begynnelsen eller slutten av opptaket.

D. Dersom bakgrunnslyder overdøver talen slik at det er vanskelig å høre hva informanten sier brukes karakteren D. D brukes også når feil uttale eller nølelyder forekommer. Dersom det er 0 ms mellom begynnelsen eller slutten av talesignalet og begynnelsen eller slutten av opptaket gis karakteren D.

E. Karakteren E fører til at opptaket blir forkastet. Dette vil være i tilfeller hvor man ikke forstår hva taleren sier, taleren leser feil ord, eller ingenting blir sagt. E brukes også hvis opptaket er avkortet.

For hvert sett av opptak er all tilhørende annotering samlet i en Speech Logging File. Dette er en ren tekstfil med filetternavn *.spl. Den inneholder fyllestgjørende metainformasjon om innspillingsutstyr, taleren, manuskriptet og de tilhørende enkeltfilene. En slik spl-fil inneholder informasjon om alle opptakene knyttet til et enkelt manuskript. Et eksempel på hva slags informasjon som finnes i en slik fil er gitt på neste side.

Eksempel på metainformasjon i spl-fil og informasjon om fire første opptak (av 320)

```
[System]
Delimiter=>-<
Version=0001_1
CharacterSet=ANSI
ByteFormat=01
Script=463
Channels=2
Board=2;NI DSP2200
Frequency=16000
Coding=PCM;Linear
DOS Codepage=850
ANSI Codepage=1252
Memo=Kontor##1,5m, 2,5m, 1,5m, 2,5m##Shure bordmikrofon##Shur

[Info states]
1=Speaker ID>-<001>-<
2=Name>-<**** ****>-<
3=Age>-<57>-<
4=Sex>-<Female>-<
5=Region of Birth>-<Voss og omland>-<
6=Region of Youth>-<Voss og omland>-<
7=Remarks>-< frå hardanger>-<

[Session]
Directory=c:\adb_0463\data\scr0463\10\04631001\r4630001
Imported sheet file=c:\adb\dsdr\scripts\nor463\nor463.psh
Record session=1
Sheet number=1
RecDate=02 jul 1999
RecTime=08:52:13
Record duration=75' 36"
Number of recordings=312

[Record states]
1=2>-<>-<(...Vær stille under dette opptaket...)>-<1024>-<257024>-
<u0001001.wav>-<>-<1024>-<257024>-<bISa1>-<bISa1
2=2>-<>-<Tester en to tre fire fem seks sju åtte>-<1024>-<561024>-
<u0001002.wav>-<>-<257024>-<817024>-<tISa1>-<tISa1
3=2>-<>-<Blåbærturen ute på landet var en rein fornøyelse og flere av
turgåerene hadde kilosvis med bær.>-<1024>-<593024>-<u0001003.wav>-
<>-<817024>-<1409024>-<cISa1>-<cISa1
4=2>-<>-<Piloten hadde sitt svare strev med å få landet flyet i uvær
og svart natt.>-<1024>-<529024>-<u0001004.wav>-<>-<1409024>-
<1937024>-<cISa2>-<cISa2
5=2>-<>-<Det kan virke helt overveldende ute ved havet når den salte
skumsprøyten slår innover holmer og skjær.>-<1024>-<577024>-
<u0001005.wav>-<>-<1937024>-<2513024>-<cISa3>-<cISa3
```

”****” angir anonymisering foretatt i denne rapporten.

5. Dialektområder

I det følgende gis en oversikt over inndelingen i dialektområder som er brukt ved oppbygging av NSTs akustiske database. Denne ligger til grunn for datainnsamling av alle deldatabasene beskrevet ovenfor. Talerne fordeler seg på aldersgruppene 18-70, og begge kjønn er representert. Det er ikke funnet noen samlet statistikk som angir fordelingen innenfor disse gruppene, men sporadiske funn i arkiverte rapporter antyder at fordelingen er noenlunde jevn. Dette bekreftes også av Kolbjørn Slethei ved Seksjon for Humanistisk Informatikk ved UiB, som tok del i ut-

arbeidelsen av dialektinndelingen. For øvrig vil informasjon om informantfordeling uansett kunne la seg ekserpere fra spl-filene.

Den norske databasen er fordelt på talere fra følgende 11 dialektområder:

- Hedmark og Oppland
- Oslo-området
- Ytre Oslofjord
- Sørlandet
- Sør-Vestlandet
- Bergen og Ytre Vestland
- Voss og omland
- Sunnmøre
- Trøndelag
- Nordland
- Troms

Det foreligger ikke dokumentasjon som angir kriteriene for denne inndelingen, men Kolbjørn Slethei opplyser at denne er hovedsakelig basert på språklige vurderinger og sekundært på statistiske og sosioøkonomiske forhold. Et direktiv fra L&H tilsa at det maksimale antall dialektområder kunne være fem, så NST måtte "forhandle" seg frem til et høyere antall dialekter og dermed høyere antall informanter enn ved tilsvarende innsamlinger for andre europeiske språk.

6. Språklige vurderinger

I det følgende kommenteres en del språklige vurderinger og veivalg som er blitt gjort under arbeidet med den norske del av databasen.

For norsk (og svensk) gjelder at det opereres med et antall dialektområder og et antall informanter som relativt sett ligger høyere enn for tilsvarende innsamlingsprosjekt på øvrige språk i L&Hs portefølje. Dette har NST måttet rettferdiggjøre overfor sin teknologipartner med bakgrunn i den store dialektvariasjonen som forekommer i norsk.

Man har bevisst valgt å fokusere på bokmål og utelate nynorsk. Dette gir seg utslag i at alle innspillingsskript kun inneholder tekster på bokmål. Dersom taleren leser et bokmålsord som nynorsk, for eksempel hvis *skole* uttales som *skule* eller *åttende* uttales *åttande*, er ordets ortografi endret under valideringen slik at annotertingen stemmer overens med uttalen. I slike tilfeller er ordet merket som nynorsk i annotasjonsfilens kommentarfelt. Dette betraktes ikke som feil, og slike tilfeller er validert i henhold til kvalitetskriterium A, forutsatt at den ortografiske formen er innenfor nynorsk skriftnorm. En liknende strategi er valgt i tilfeller hvor en dialektform erstatter en bokmålsform, som når *venner* blir uttalt *venna*. Hvis dialektformen er relativt vanlig og ikke skiller seg særlig fra skriftnormen, vil kvalitetskriterium B være benyttet.

En forutsetning for arbeid med akustiske databaser er et tilhørende uttaleleksikon med informasjon om ordenes ortografi og uttale. En beskrivelse av NSTs leksikalske databaser ligger på Språkbankens nettside, og databasene kan lastes ned derfra. Dette leksikonet gjenspeiler uttaler som forekommer i den akustiske databasen. Det er tatt høyde for en viss grad av fonetisk variasjon i den akustiske databasen under utarbeidelsen av uttaleleksikonet. Leksikonet inneholder uttalevarianter i de tilfeller hvor de forekommer naturlig i opptakene og i talemål gene-

relt; eksempelvis *vende* som [ˀvɛ.nə]/[ˀvɛn.də], og flere aksepterte uttaler av *morgen*, *måned*, *tredje*, *sytti* osv. Talerens faktiske uttale er ikke angitt i fonetisk representasjon i spl-filen men den er representert i leksikonet.

Fonetisk reduksjon er håndtert på ulikt vis, avhengig av konteksten den forekommer i. En del reduserte uttalevarianter vil få kvalitetsmerking A; dette gjelder varianter hvor reduksjon så å si alltid forekommer, som i *meteorolog* og *amerikaner* uttalt som henholdsvis [ˌmɛ.tɾʊ.ˈloːg] og [ˌɑm.rɪ.ˈkɑː.nəɾ]. Dette regnes som den riktige uttalen av disse ordene, og i slike tilfeller vil faktisk den mer ortofone varianten betraktes som overartikulert og dermed få kvalitetsmerke B. Kvalitetsmerke B grunnet overartikulering gis også ved unaturlige geminater (*telefonnummer* uttalt med to *n*-er), eller ved bestemte nøytrumsformer hvor endelsen blir artikulert med plosiv (*hodet* uttalt [ˀhuːdət]). Ved ikke-obligatoriske, men vanlige, reduksjoner vil kvalitetskriterium B kunne være benyttet; eksempelvis *forutsette* uttalt [ˈfɔɾ.ʉ.ˌsɛ.tə] eller *Sarpsborg* uttalt uten [p]. Dette gjelder også f.eks. når *Hurdal* uttales [ˈhʉ.ˌɖɑɳ], hvor også annoteringens ortografi er endret til *Hurdalen*, i samsvar med uttalen.

Ved sjeldnere og uønskete reduksjoner brukes kvalitetsmerking fra C til E, avhengig av grad. Eksempelvis vil dialektformene *saukjan*, *akjan* og *nikkjan* få kvalitetsmerking E, og dermed bli forkastet som realisasjoner av tallordene.

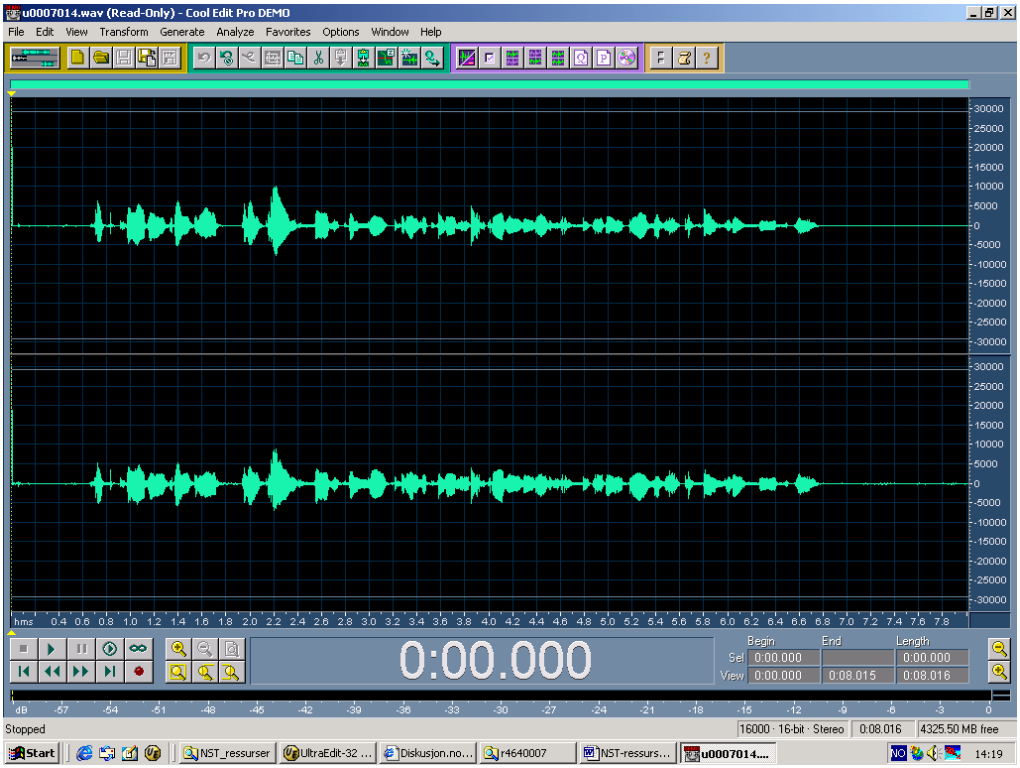
Fonetisk/allofonisk variasjon forekommer naturlig i dataene og er annotert etter tilsvarende kvalitetskriterier. I de fleste tilfeller vil et sett av varianter være akseptert etter kvalitetskriterium A. Dette gjelder for eksempel variasjon av typen *kino* /çiː.nʊ/ vs. /ʃiː.nʊ/, *kort* /ˈkɔɯt/ vs. /ˈkɔʈ/ og *opp* /ˈɔp/ vs. /ˈʊp/. Normalt vil variasjonen være representert i form av fonemisk ulike varianter i uttaleleksikonet, men ikke i tilfellet *kino*, som kun er representert ved én uttaleform symbolisert ved [ç]. Konsekvensen av dette er at de to realisasjonene inngår som allofoniske varianter av fonemet [ç] i språkmodellene.

I mer dialektspesifikke tilfeller vil varianter kunne være merket med andre kriterier enn A. For eksempel er /b d ɡ/ for [p t k] (bløte konsonanter) i sørlandsområdet annotert med kvalitetskriterium B, mens trøndersk /ʃ/ i verbformen *ser* ikke er godkjent, men kvalitetsmerket som E.

Det er grunn til å merke seg at databasen ikke er spesifikt konstruert for å representere fremmedspråklig uttale, men det kan forekomme informanter med utenlandsk opprinnelse i materialet. Instruksjonene er på dette feltet ikke entydige, men det later til at fremmedspråklig uttale godtas som norske uttalevarianter dersom den følger et konsekvent mønster, for eksempel bruk av /iː/ for [yː] (*lyd*) eller /uː/ for [ʉː] (*du*). Denne uttalen vil imidlertid ikke være merket med kvalitetskriterium A.

7. Kvalitetsvurdering

Stikkprøvekontrollen som er foretatt som del av gjeldende undersøkelse tyder på at lydmaterialet er av generelt høy kvalitet og grundig og nøyaktig validert. Lydkvaliteten er gjennomgående utmerket, med minimale mengder støy. Opptakene vil bli automatisk kuttet av innspillingsprogramvaren dersom talen overstiger et visst lydnivå. Ved stikkprøvekontrollen er det ikke funnet opptak hvor dette forekommer. Dette sees blant annet ved spektrogramvisning, hvor eventuelle kuttete opptak vil ha en flat kurve øverst, i stedet for å ha naturlige topper (jf. figurene under).



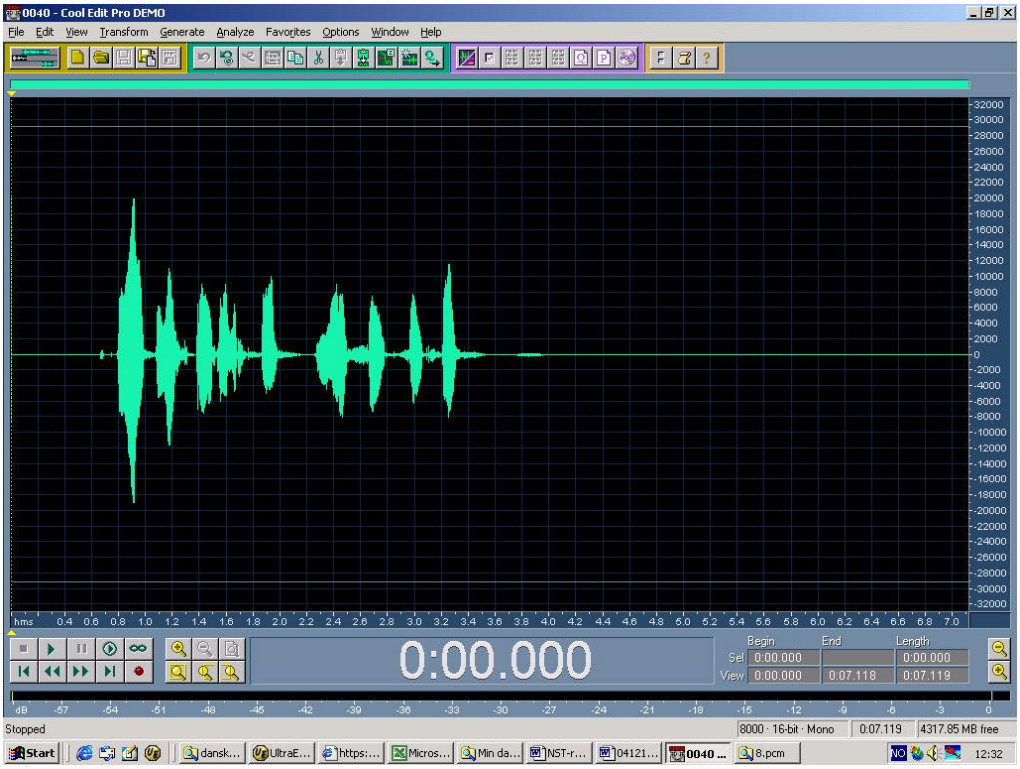
Eksempel på stereoopptak (8 kHz, kontorbasert)

Som nevnt har man under innspillingen lagt inn 100-200 millisekunders tomrom mellom begynnelsen og slutten på talesignalet og begynnelsen og slutten på opptaket. Ved stikkprøvekontroll ble det funnet noen få telefoniopptak som var klippet på slutten, hvor siste del av setningen var utelatt. Problemets omfang er ikke klart, men det synes lite. Det er ikke funnet noen klippete opptak i opptakene fra kontormiljø. Uansett vil slike klippete opptak være fjernet ved hjelp av annotering under valideringen.

Kvalitetssikring av dataene gjennom validering later til å være grundig gjennomført. Valideringen fremstår som konsistent, noe som må tilskrives grundig dokumentasjon og konsekvent veiledning av språkassistenter. Informantenes tale er tydelig annotert ved hjelp av markører, den og lar seg skille fra uønsket lyd som innpust, utpust, spyttlyder, nøleøyder og annen støy. Dette gjør det mulig å produsere akustiske modeller som ikke inneholder støyforstyrrelser. Valideringen inneholder som nevnt også detaljerte opplysninger om eventuelle feiluttaler som talerne gjør.

Det må påtales at knirkestemme (*creaky voice*) forekommer i dataene uten at det er eksplisitt merket eller fjernet ved annotering. Dette er imidlertid ikke nødvendigvis noen svakhet, men kan inngå som en del av modelleringen da det også forekommer i naturlig tale. Under opptak for talesyntese er segmenter med knirkestemme fjernet eller overspilt.

Foruten fullstendig validering av dataene er det foretatt systematiske stikkprøvekontroller, først av NSTs gruppeledere og deretter av L&Hs personale. Om lag 10 prosent av materialet er stikkprøvekontrollert av NST, mens 5 prosent er stikkprøvekontrollert i Belgia etter leveranse. Denne manuelle kontrollen overgår dermed de fem prosentene som ELRA-standarden krever. Apropos kvalitet og spotsjekking, i en intern møterapport fremgår følgende:



Eksempel på monoopptak (8 kHz, telefoni)

NN fortalte først litt om situasjonen for dikteringsvalideringen. NST får ros for at vi har mye data, dvs. 900 opptak som er ferdigvalidert. I Belgia har de spotsjekket 150 dikteringsopptak og det er meget god kvalitet på disse. Derfor er det besluttet i Belgia at de dropper sin videre "interne" spotsjekking, …

Ressursene fremstår som godt dokumenterte, og det er konsistens i navngiving av filer og katalogstruktur.

Det må fremheves at valideringsgruppene har hatt hyppige møter og har vært under faglig koordinering av lingvistisk skolerte gruppeledere. Valideringsarbeidet har vært samordnet mellom de tre språkene, og språkressursene har blitt bygget opp parallelt. Jevnlige møter innenfor og på tvers av de språkspesifikke gruppene gjør det klart at har NST gjort en betydelig innsats for å samordne arbeidet og utvikle en felles standard for valideringen. Dette har gitt materialet et entydig preg.

Det må imidlertid påpekes at en del av dokumentasjonen foreligger kun på norsk, noe som vil måtte endres dersom materialet skal inngå i en internasjonal ressursdatabase (jf. ELRAs kravspesifikasjon).

Oppbygging av akustiske databaser er ressurskrevende, og NST har lagt ned en stor innsats i oppgaven med å bygge opp en akustisk database bestående av i alt nærmere 2 millioner opptak. Bare i planleggingsfasen krever en slik ressursoppbygging omfattende oppgaver, som kartlegging av dialektområder, skriptkonstruksjon, prosjektstyring, ansettelser av personale til innsamlingen, opplæring, innkjøp av opptaksutstyr, transport av utstyr og personell til innspillingslokasjoner, rekruttering av informanter, kontrakter, logistikk osv. I neste fase kommer selve arbeidet med opptakene, før valideringen av foretas.

En naturlig innvending er at inndelingen i dialektområder ikke helt og holdent er gjort i henhold til dokumenterte dialektskiller, slik disse er beskrevet i litteraturen. Det virker som man har tatt praktiske og markedsmessige hensyn, mer enn hensyn til reelle dialektskiller i utvelgelsen av områder for innsamling av data. Man kan spørre seg hvorfor Voss er representert mens andre regioner med særpregete dialekter, som Sogn, Sunnfjord osv. ikke er tatt med. Det er imidlertid vanskelig å vurdere eventuelle konsekvenser dette valget har på reell gjenkjenningsrate, og om det er et mål å representere hver eneste dialekt i et så komplekst område som det nordiske. Viktigere er det å slå fast at et bredt geografisk område er representert i databasene for de tre språkene.

En annen betimelig innvending, som gjelder kun for norsk, er at nynorsk ikke er ivaretatt. Begge hovedmanuskriptene inneholder kun bokmålstekst. Det er presumptivt markedsmessige hensyn som ligger til grunn her, og språkpolitisk kan dette være problematisk. Riktignok representerer innsamlingen et stort geografisk område som inkluderer typiske nynorskområder, men det kan være fremtidig behov for å supplere materialet med nynorskbaserte opptak. For øvrig har NST valgt å håndtere forekomster av nynorskformer i den akustiske databasen på en forsvarlig måte.

Tross disse innvendingene kan det konkluderes med at NSTs akustiske database er i henhold til gjeldende standard slik denne er formulert etter ELRA-standarden, når det gjelder krav til teknisk kvalitet, språklig kvalitet, grad av validering, stikkprøvekontroll, konsekvente metoder og dokumentasjon.

8. Akustiske databaser for talesyntese

Deldatabasene beskrevet under kan lastes ned via følgende lenke:

- http://www.nb.no/sbfil/talesyntese/no.ibm.talesyntese.tar.gz

NST har i flere omganger gjort opptak for produksjon av talesyntese, først i forbindelse med utvikling av skandinaviskspråklige versjoner av L&H-programvaren RealSpeak, som fremdeles er tilgjengelig fra selskapet Nuance, deretter for produksjon av IBMs talesyntese, som ikke rakk å nå markedet før NST gikk konkurs. Begge syntesene er konkatenative systemer; førstnevnte er en difonsyntese, mens IBM-syntesen er en datadrevet skjøtesyntese (unit-selection synthesis). I tillegg har NST utviklet IBM-baserte testsystemer for domenespesifikk syntese, under benevnelsen Phrase Splicing (frasespleising), samt gjort opptak for en del spesifikke kundeapplikasjoner.

8.1. RealSpeak

Opptakene som ble gjort for utvikling av RealSpeak foregikk i sin helhet i L&Hs innspillingsstudio i Ieper, Belgia. Disse opptakene er ikke en del av dataene i konkursboet etter NST.

8.2. IBMs talesyntese

I forbindelse med utviklingen av IBMs talesyntese ble det rekruttert profesjonelle stemmer, dvs. én herrestemme per språk. Opptakene er innspilt med IBM-programvare i et lydstudio på Voss,

men proprietær innspillingsprogramvare er ikke til hinder for fremtidig bruk, da opptakene foreligger i anvendelig PCM-format. Følgende teknisk informasjon gjelder for de tre deldatabasene:

Signalkoding:	lineær PCM
Filformat:	ukodete rådata (headerless raw)
Samplingsfrekvens:	44 kHz
Oppløsning:	16 bit
Format:	Motorola PCM
Kanaler:	2 (stereo): tale + laryngograf

Stereoopptakene har talesignal i én kanal, og signal fra laryngograf i andre kanal. Innspillingsmanuskriptene er basert på NSTs korpus. Et optimalisert utvalg av setninger er produsert ved hjelp av IBMs programvare OptScript. Manuskriptene er fonemisk optimaliserte for å oppnå bred dekning av mulige difonkombinasjoner. Manuskriptene er ikke prosodisk balanserte, men er likevel fordelt på kategorier som innebærer en viss prosodisk variasjon, som fortellende setninger, *hv*-spørsmål *ja/nei*-spørsmål og opplistinger. Et mindre subsett av manuskriptene inneholder setninger og tallformater som trengs i forbindelse med utvikling av en spesialisert taleapplikasjon for bankdomenet.

Ressursene fordeler seg som følger:

Antall opptak:	5363
Hvorav bank:	417

8.3. IBM Phrase Splicing

I tillegg til ovennevnte opptak beregnet på kommersiell utnyttelse ble det spilt inn flere datasett for produksjon av test- og demonstrasjonssystemer av IBMs programvare for frasespleising (Phrase Splicing), et system som er en hybrid mellom applikasjonsspesifikke innspillinger og vanlig konkatenativ talesyntese. Systemene er beregnet for bruk innenfor bankdomenet.

Følgende teknisk informasjon gjelder for disse deldatabasene:

Signalkoding:	lineær PCM
Filformat:	ukodete rådata (headerless raw)
Samplingsfrekvens:	44 kHz
Oppløsning:	16 bit
Format:	Motorola PCM
Kanaler:	2 (stereo): tale + laryngograf

Dette er de samme tekniske spesifikasjoner som for dataene beskrevet i avsnitt 8.2.

Imidlertid er det en vesensforskjell når det gjelder materialets omfang og kvalitet. Disse opptakene er ikke gjort av profesjonelle stemmer, men informantene er i sin helhet rekruttert blant NSTs egne ansatte. Manuskriptene og antall opptak er også mindre enn for datasettene nevnt ovenfor.

8.4. Kvalitetsvurdering

Opptakene gjort for utvikling av IBMs talesyntese, nevnt i avsnitt 8.2, er gjort i lydstudio og med opptaktutstyr og laryngograf som er godkjent av IBM for produksjon av kommersielle talesyntesesystemer. Det er grunn til å understreke den høye samplingsfrekvensen, og den generelt høye kvaliteten på selve opptakene, samt at informantene er profesjonelle aktører. Opptakene har vært manuelt kontrollerte av NSTs produktutviklere for talesyntese. Disse akustiske databasene er segmentert og merket ved hjelp av IBMs annoteringsverktøy. Annoteringen er først gjort maskinelt og deretter manuelt kontrollert av utviklerne.

Materialet er i henhold til gjeldende standard for akustiske ressurser, og representerer ”state-of-the-art” for utvikling av talesyntese. Det er dermed anbefalt at inngår som del av en videreføring av NSTs akustiske database. Det vil også være svært gunstig dersom man fikk tilgjengeliggjort annoteringene i forbindelse med segmentering og parallellstilling (alignment) av lyd og tekst. Dette forutsetter antakelig en avtale med IBM på dette punktet.

Opptakene til frasespleising nevnt i avsnitt 8.3 er gjort som del av opplæring av NSTs ansatte i utvikling av syntseteknologi på IBMs avdelinger i Heidelberg, Hursley og Paris. Opptakene er gjort i til dels støyfulle kontormiljøer, og ikke i lydstudio. Dette gjør dem lite egnet til fremtidig bruk. For øvrig har disse databasene ikke samme fonetiske dekningsgrad som de som er omtalt i avsnitt 8.2.