Modellering av uttalevariasjon for automatisk talegjenkjenning

Ingunn Amdal, Institutt for teleteknikk, NTNU/Telenor FoU
Trym Holter, SINTEF Tele og data
Torbjørn Svendsen, Institutt for teleteknikk, NTNU

Sammendrag

I denne artikkelen tar vi for oss problemet med å optimalisere uttale-leksikonet for bruk i taleruavhengig automatisk talegjenkjenning. Vi bruker talte eksempler av ordene i vokabularet for å finne uttaleformer som tilfredsstiller optimaliseringskriteriet "maximum likelihood" (ML). Den ML-baserte algoritmen vi foreslår kan finne flere uttaleformer for hvert ord og på den måten modellere uttalevariasjon, [1, 2]. Eksperimenter med tale fra 200 talere fra hele landet viser at disse automatisk genererte uttaleformene gir bedre gjenkjenningsrate enn de originale, manuelt genererte uttaleformene.

1. Innledning

Modellering av uttalevariasjon er et viktig tema for forskning innen talegjenkjenning for tida. Målet er å lage brukervennlige systemer basert på taleteknologi. For å unngå å legge unødvendige restriksjoner på brukerne, er det viktig å kunne håndtere variasjon mellom talerne. Noe av variasjonen mellom forskjellige talere vil typisk håndteres av de akustiske modellene (se kapittel 2), men en del variasjon vil være felles for grupper av talere (for eksempel dialekt, ordvalg og talerate) og kan bedre

modelleres eksplisitt i uttaleordlisten. Det er dette vi tar for oss i denne artikkelen.

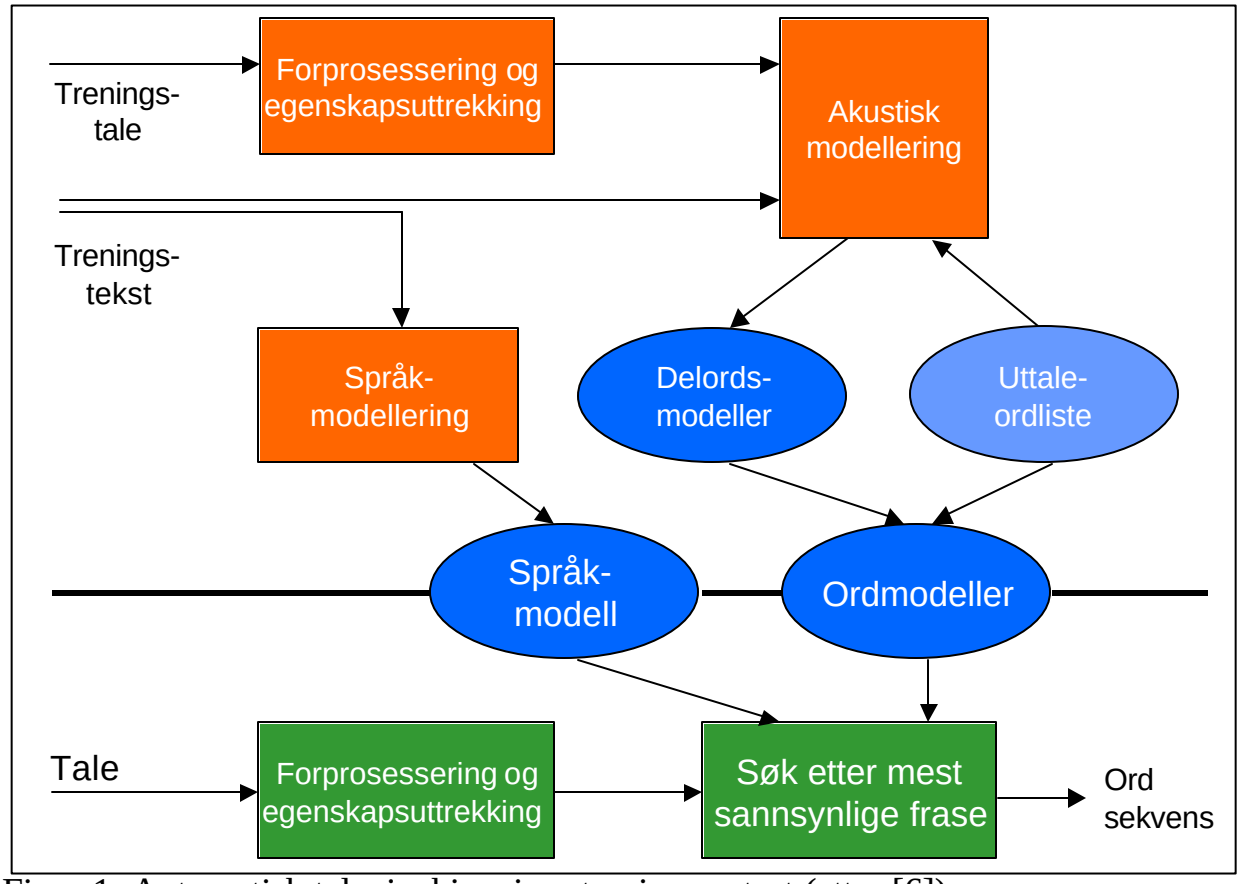
Vokabularet som er valgt i eksperimentene er norske naturlige tall (0–99), og tale-eksemplene er tatt opp over telefonlinjer. Tidligere eksperimenter på samme oppgave har vist behovet for modellering av uttalevariasjon, [3, 4]. Tallordene er korte og like. De er samtidig hyppig brukte ord, og vil derfor ofte uttales med mer variasjon enn andre ord, [5].

2. Automatisk talegjenkjenning

Et talegjenkjenningssystem kan grovt sett deles i tre deler, se figur 1:

- Språkmodellen er en statistisk beskrivelse av sannsynligheten for at forskjellige ordsekvenser inntreffer.
- De akustiske modellene er statistiske beskrivelser av talesignalet for hvert enkelt delord (som oftest fonem/fon-liknende enheter).
- Uttaleordlisten (ofte kalt leksikon) beskriver hvordan delordsmodellene settes sammen til ordmodeller.

For trening av både språkmodellen og de akustiske modellene, trenger vi treningsdata, henholdsvis i form av tekst og transkribert tale. Det er viktig at disse dataene inneholder den variasjonen vi ønsker at modellene skal kunne ta opp i seg. I de aller fleste tilfellene trenes modellene med hensyn på ML-kriteriet. Dette innebærer å finne de statistiske modellene som best beskriver (eller forklarer) treningsdataene. De ferdig trente matematiske modellene utgjør sammen med uttaleordlisten de språkspesifikke komponentene i en automatisk talegjenkjenner.



Figur 1: Automatisk talegjenkjenning: trening og test (etter [6]).

3. Data-drevet uttalemodellering

Uttaleordlisten står i en særstilling av de språkspesifikke komponentene i et talegjenkjenningssystem, siden den i motsetning til språkmodellen og de akustiske modellene normalt ikke optimaliseres ved bruk av treningsdata. I dette arbeidet beskriver vi en metode for å optimalisere ordlista ved bruk av treningsdata. Denne metoden baserer seg også på ML-kriteriet, og vi får dermed et konsistent rammeverk for trening.

Som oftest velger vi å modellere delord i form av fon/fonem-liknende enheter. Vi bruker begrepet baseformer om uttaleformer beskrevet ved hjelp av gjenkjennerens akustiske modeller. Som en følge av unøyaktigheter og artefakter i treningen, er det ikke nødvendigvis en

presis sammenheng mellom disse modellene og de underliggende enhetene. Dette betyr at den fonetiske og fonologiske kunnskap vi har om uttale av ord ikke nødvendigvis resulterer i optimale baseformer. I tillegg har vi ofte ikke tilstrekkelig fonologisk kunnskap, særlig ikke i forbindelse med dialekt og sjargong. Det er også sannsynlig at brukere av et automatisk talegjenkjenningssystem vil bruke andre uttaleformer enn de ville gjort hvis de snakket med et annet menneske. Alt dette peker i retning av at data-drevet modellering av uttalevariasjon kan være veien å gå, men vi må også ta hensyn til hvor mange som bruker de forskjellige uttaleformene. Dersom alle tenkelige uttaleformer innføres i leksikonet, fører dette både til økt kompleksitet og redusert nøyaktighet i gjenkjenneren fordi mulighetene for forveksling øker.

Den vanligste måten å modellere uttalevariasjon på, er å øke antall baseformer pr. ord. En kan for eksempel bruke kunnskap om uttalevariasjon til å legge inn flere uttaleformer eller generere varianter ved å bruke regler for uttalevariasjon (enten med utgangspunkt i eksisterende uttaleformer eller grafem-til-fonem oversettere). Igjen risikerer vi å ende opp med mange like baseformer og dermed øke faren for forveksling av ord.

I algoritmen som presenteres her, brukes ML-kriteriet for både å velge antall baseformer for hvert ord og hvilke baseformer dette skal være. Utvelgelsen gjøres med utgangspunkt i de akustiske modellene og et treningssett med talte eksempler. Målet er å finne det riktige antall baseformer for hvert ord og de riktige formene. Vi vil at ord med liten uttalevariasjon skal ha få baseformer og ord med stor uttalevariasjon skal ha flere baseformer.

I disse eksperimentene har vi ikke lagt noen restriksjoner på baseformene, det vil si at alle mulige delordsekvenser av enhver lengde tillates. I en videreføring vil det være aktuelt å legge inn fonotaktisk kunnskap, men i disse eksperimentene står gjenkjenneren fritt til å velge baseformer kun på grunnlag av de akustiske talesignalene. I valget av baseformer ønsker vi at hver baseform skal representere en gruppe av talere eller uttaler, slik at det er variasjon mellom distinkte uttaleformer vi modellerer. Dette gjøres ved å bruke en algoritme der vi automatisk deler taleeksemplene inn i klynger. Hver klynge representerer en uttaleform, og vi tilordner en baseform til hver av disse. Til slutt velger vi antall baseformer for hvert av ordene ved å sette en grense for antall baseformer totalt i leksikonet og så bruke ML-kriteriet til å finne den beste kombinasjonen.

4. Eksperimentene

Vi brukte norske naturlige tall (0–99) i eksperimentene. For å kunne gjenkjenne disse trenger vi 29 ord, se tabell 1. I tillegg til at tallordene er korte og like og dermed vanskelige å gjenkjenne, er det lettere å analysere feil med et såpass lite vokabular. Opptakene vi brukte er fra to databaser med telefontale tatt opp ved Telenor FoU; TABU.0, [7], og SpeechDat, [8]. SpeechDat er en del av et europeisk prosjekt hvor liknende databaser er samlet inn for alle europeiske språk.

I begge disse databasene er det ca. 1000 talere av begge kjønn og alle aldre (ikke små barn). Talerne leste fra et manuskript, og denne opplesingen ble tatt opp over telefon. For tallene var manuskriptet oppgitt med tall slik at talerne selv valgte om de for eksempel ville si “sju” eller “syv” for 7.

Ord	Ortografisk transkr.	Fonemisk transkr.	Ord	Ortografisk transkr.	Fonemisk transkr.
0	null	/n } l/	17	sytten	/s 2 t n/
1	en	/e: n/			/s y t n/
	ein	/{i n/			/s 2 t e n/
2	to	/t u:/			/s y t e n/
3	tre	/t r e:/	18	atten	/A t n/
4	fire	/f i: r e/			/A t e n/
5	fem	/f e m/	19	nitten	/n i t n/
6	seks	/s e k s/			/n i t e n/
7	sju	/S }:/	20	tjue	/C }: e/
	syv	/s y: v/		tyve	/t y: v e/
8	åtte	/O t e/	30	tretti	/t r e t i/
9	ni	/n i:/		tredve	/t r e d v e/
10	ti	/t i:/	40	førti	/f 2 rt i/
11	elleve	/e l v e/		førr	/f 2 r/
12	tolv	/t O l/	50	femti	/f e m t i/
13	tretten	/t r e t n/	60	seksti	/s e k s t i/
		/t r e t e n/	70	sytti	/s 2 t i/
14	fjorten	/f j u rt rn/			/s y t i/
		/f j u rt e n/	80	åtti	/O t i/
15	femten	/f e m t n/	90	nitti	/n i t i/
		/f e m t e n/		og	/O:/
16	seksten	/s {i s t n/			
		/s {i s n/			
		/s e k s t n/			
		/s {i s t e n/			

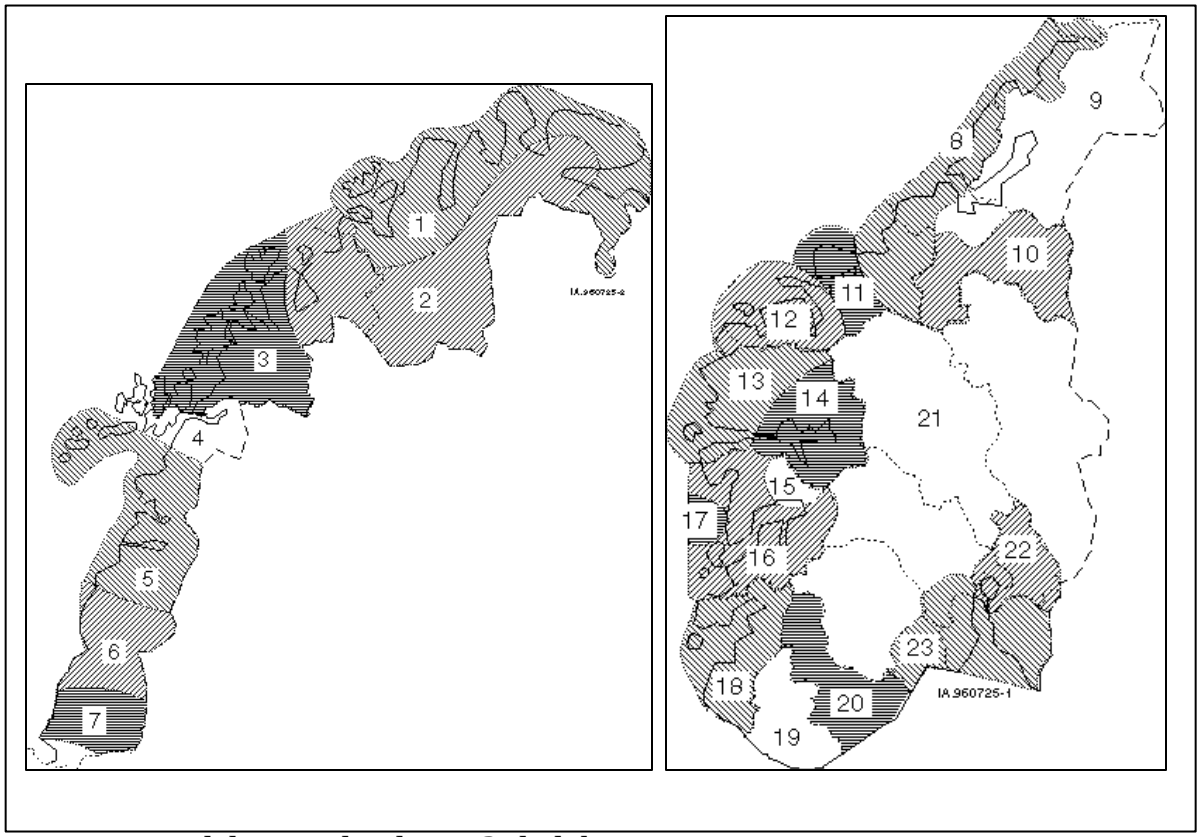
Tabell 1: De 29 ordene i norske naturlige tall 0-99 med de 46 uttaleformene fra SpeechDat, fonemisk transkripsjon i SAMPA, [9].

Vi brukte 200 ytringer av hvert ord fra 382 av talerne i TABU.0 til å finne det nye leksikonet. 200 andre talere ble brukt til testsett, de leste til sammen 13821 ord i 2330 tallstrenger. Tallstrengene var i utgangspunktet 8-sifrede telefonnummer, men feil i opplesingen har ført til både lengre og kortere tallstrenger.

For hver taler ble dialekt-tilhørighet registrert ved å velge en av 23 dialektregioner, se figur 2. Fordelingen av antall talere i hver region var omtrent tilsvarende folketallet, med en viss overrepresentasjon i mindre tettbygde strøk. Test-talerne fordelte seg dermed med fra 4 til 28 talere i

hver region. For å få et bedre grunnlag for analysen delte vi de 23 regionene inn i 5 større regioner; nord, midt, vest, sørvest og sørøst.

De akustiske modellene som vi bruker er trent på SpeechDat-databasen, og referanse-leksikonet vårt er også tatt derifra. Som vist i tabell 1 er det lagt inn 46 uttaleformer for de 29 ordene i tall-vokabularet her. Dette er adskillig flere enn i for eksempel NorKompleks som har 32 uttaleformer (kun sørøstnorsk). Vi antar derfor at dette referanse-leksikonet bør dekke mange av de uttaleformene vi vil finne.



Figur 2: Inndeling av landet i 23 dialektregioner.

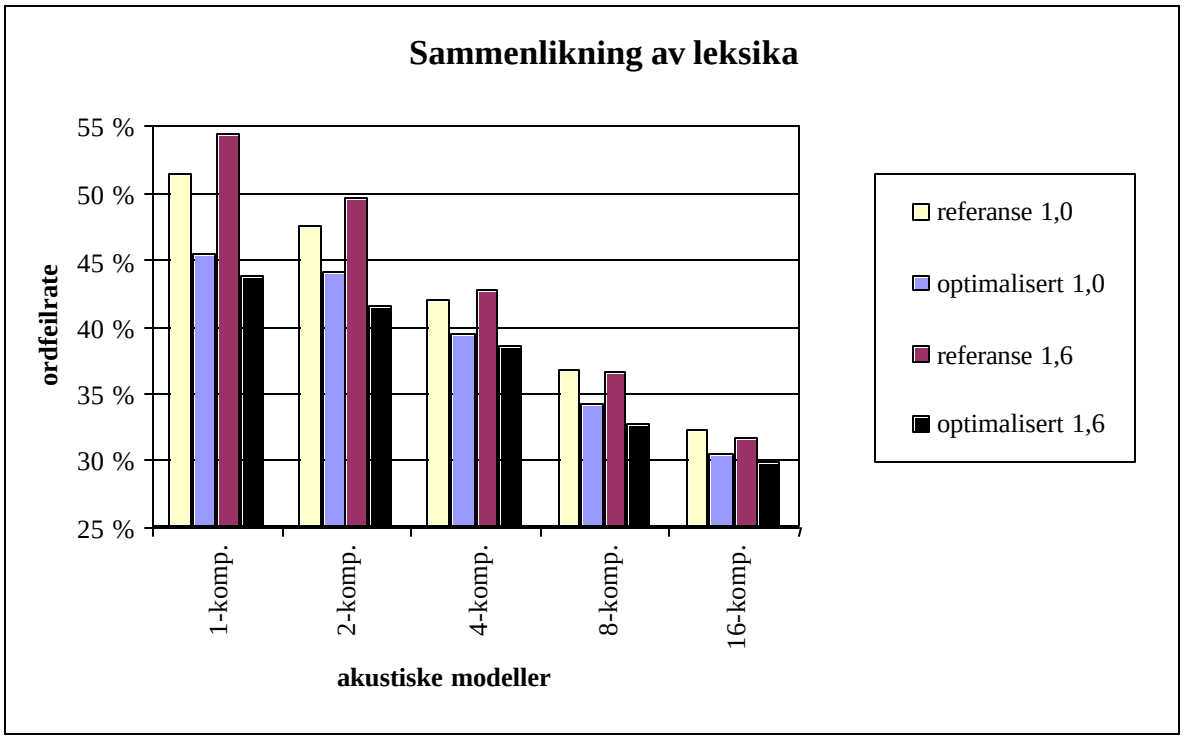
5. Resultater

Figur 3 viser en oversikt over resultatene. Gjenkjennerens kompleksitet er oppgitt i antall komponenter i sannsynlighetstetthetsfordelingene som

inngår i de akustiske modellene, for eksempel er 16-komponents modeller mer komplekse enn 1-komponents modeller og vil kunne modellere det akustiske signalet mer detaljert. De akustiske modellene som er angitt er brukt både i optimalisering av leksikonet og i test.

Vi ser at ytelsen øker (ordfeilraten blir mindre) med mer komplekse akustiske modeller både for referanse-leksikonet og det optimaliserte ML-leksikonet. 1,0 og 1,6 refererer til antall baseformer i snitt pr. ord. 1,0 betyr da at vi kun har brukt en baseform pr. ord, for referanse-leksikonet er dette den første uttalen for hvert ord i tabell 1. Siden referanse-leksikonet har i snitt 1,6 baseformer pr. ord har vi sammenliknet med ML-leksikonet av samme størrelse. Vi eksperimenterte med fra 1,0 til 3,0 baseformer i snitt for det optimaliserte leksikonet, og resultatet for 1,6 baseformer i snitt pr. ord var ikke signifikant forskjellig fra det beste resultatet som ble oppnådd med 1,8 baseformer pr. ord.

Som vi ser av figur 3 er faktisk de ekstra uttaleformene i referanse-leksikonet ikke til hjelp for gjenkjenner-ytelsen i det hele tatt, resultatet er i flere tilfeller bedre med kun en baseform pr. ord. For ML-leksikonet oppnår vi derimot en forbedring med flere baseformer pr. ord. For mer komplekse akustiske modeller er forbedringen med ML-leksikon mindre. Det kan komme av at disse modellene har en mer presis sammenheng med fonemer og at de manuelt genererte uttaleformene derfor stemmer bedre. Det er likevel for alle eksperimentene en signifikant forbedring med ML-leksikon i forhold til referanse-leksikonet.



Figur 3: Resultater fra eksperimentene.

Ved analyse av de to leksikaene ser vi at den data-drevne prosedyren velger å tilordne to eller flere baseformer til de samme ordene som i referanse-leksikonet er modellert med mer enn en baseform. Det er altså godt samsvar mellom de objektive og subjektive kriteriene mhp. hvilke ord som har størst uttalevariasjon. Ved å gradvis øke totalt antall baseformer i ML-leksikonet , ser vi hvilke ord som er best og dårligst modellert med kun en baseform:

- Dårligst modellert med kun en baseform: 7, 13, 14, 16 og 30.
- Best modellert med kun en baseform: 9 og 10.

Tabell 2 viser ML-leksikonet hvor akustiske modeller med 16 komponenter er brukt i optimaliseringen.

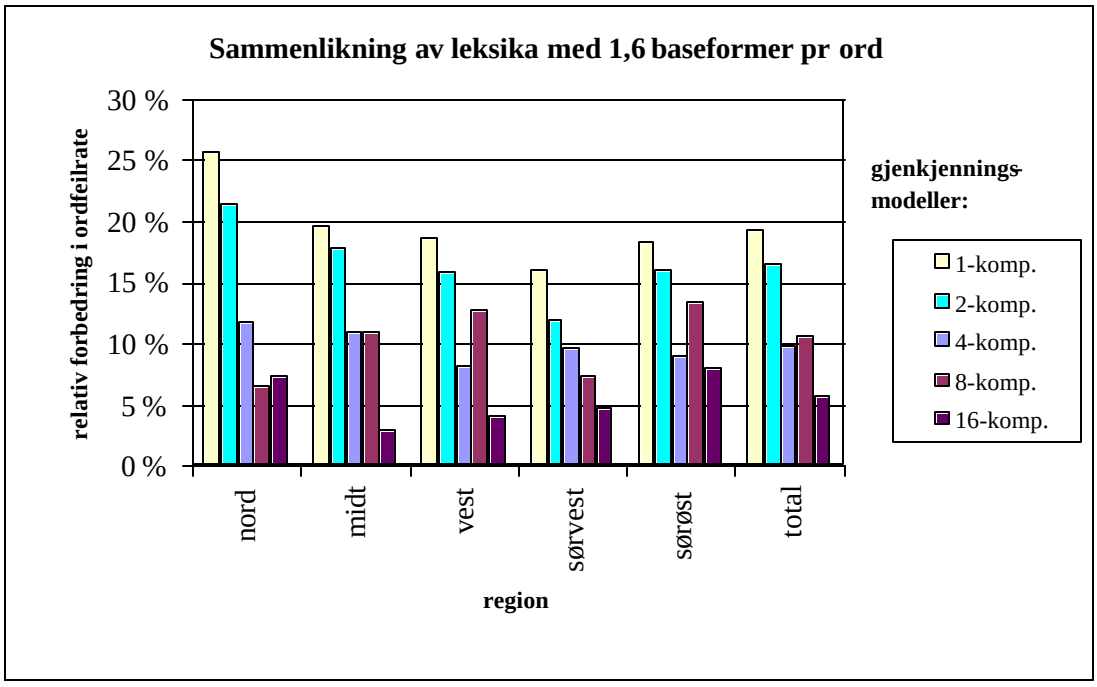
Vi sammenliknet også antall identiske baseformer i de to leksikaene og observerte at det med mer komplekse akustiske modeller var større sammenfall med referanse-leksikonet. Med 1-komponents modeller var

Ord	Ortografisk transkr.	Fonemisk transkr.	Ord	Ortografisk transkr.	Fonemisk transkr.
0	null	/n } l/	17	Sytten	/s 2 rt t } rn/
1	en	/e: m/			/s 2 rt t n/
		/e: {i n/			/s y t n/
2	to	/t u:/	18	Atten	/A rt t n/
3	tre	/t r e:/	19	Nitten	/n i t } n/
4	fire	/f i: r e/			/n i y t n/
		/f y: r 2:/	20	Tjue	/C }: e/
5	fem	/f {: m/			/C y: v 2:/
		/f e rn/	30	Tretti	/t r {: rd d v {:/
6	seks	/s e k s/			/t e t i/
7	sju	/s y: l/	40	Førti	/f 2 rt i/
		/S }:/	50	Femti	/f e m t i/
8	åtte	/O rt t e/	60	Seksti	/s e k t i/
9	ni	/n i:/			/s 2 y k s t i/
10	ti	/t i:/	70	Sytti	/s 2 t i/
11	elleve	/e l v e/	80	Åtti	/O rt t i/
12	tolv	/t O: l/			/O: t i/
13	tretten	/t e rt t n/	90	Nitti	/n i t i/
		/t e rt t } l/		Og	/O:/
14	fjorten	/f j u rt p n/			
		/C u rt t n/			
15	femten	/f e n s t } l/			
		/f A m t n/			
		/f e m t n/			
16	seksten	/s {i s t n/			
		/s e k s t } rn/			
		/f {i s t } l/			
		/f {: {i r/			

Tabell 2: Optimalisert ML-leksikon med 47 baseformer for 29 ord.

det 6 baseformer i ML-leksikonet som også fantes i referanse-leksikonet med én baseform pr. ord, og med 16-komponents modeller var det 16 slike baseformer.

Vi ville også undersøke dialekteffekter og analyserte resultatene i hver av de 5 regionene. Figur 4 viser den relative forbedringen som vi oppnår ved å erstatte det manuelt genererte leksikonet med ML-leksika optimalisert for de respektive akustiske modellsettene. Det er forskjeller i ytelse mellom de forskjellige regionene, men det er ikke noen entydig tendens



Figur 4: Resultater fra eksperimentene.

når vi ser på akustiske modeller med forskjellig kompleksitet. Akustiske modeller med lav kompleksitet modellerer detaljer dårligere enn de mer komplekse modellene. Det kan derfor være at ML-leksikonet i dette tilfellet kompenserer for dette ved å modellere allofoniske varianter, vi ser at det særlig er region nord som har forbedring for de minst komplekse modellene. Ved nærmere inspeksjon kan vi også se eksempler på mer eller mindre allofoniske varianter i ML-leksikonet basert på 1-komponents modeller (i SAMPA-notasjon, [9]):

- /e: rn n/ og /{: {i m/ for 1
- /s 2 r t i/ og /s 2 y t i/ for 70

I ML-leksikonet basert på 16-komponents modeller finner vi eksempler på at baseformene modellerer mer fonemisk variasjon:

- /s {i s t n/ og /s e k s t } rn/ for 16
- /C }: e/ og /C y: v 2:/ for 20

Vi observerte også ei anna fordeling av uttalene innafor hvert ord med ML-leksikon, et eksempel er to baseformer for 5 hvor /f {: m/ ble valgt i 80-90% av forekomstene i regionene nord, midt og vest, men bare i 50% av forekomstene i sørøst-regionen. Her var baseformen /f e rn/ hyppigst valgt.

6. Konklusjon

I denne artikkelen har vi beskrevet en algoritme for å lage uttaleleksikon for bruk i automatisk talegjenkjenning ved hjelp av objektive kriterier og et treningssett. Eksperimentene viser en signifikant bedre ytelse med slike leksika, sammenliknet med referanse-leksikonet:

- Forbedringen med optimalisert leksikon i forhold til referanse-leksikonet er størst i de tilfellene hvor vi bruker flere baseformer pr. ord. Det tyder på at den data-drevne algoritmen er minst like godt egnet til å identifisere relevante baseformer for modellering av uttalevariasjon, sammenliknet med de tradisjonelle metodene.
- Det er et visst sammenfall av baseformer når en sammenlikner ML-leksikonet og referanse-leksikonet der begge har en baseform for hvert ord. Det blir ikke større sammenfall ved å øke antall baseformer i leksikonet.
- Det er vanskelig å se klare dialekteffekter ved oppdeling av landet i 5 regioner.
- Med mer komplekse akustiske modeller reduseres gevinsten ved bruk av ML-leksika.
- Det er større sammenfall mellom ML-leksika og referanse-leksika når det benyttes mer komplekse akustiske modeller.

7. Videre arbeid

De akustiske modellene som er brukt i disse eksperimentene gir ikke god nok ytelse til bruk i et praktisk system. Vi har i første rekke ønsket å undersøke potensialet til den foreslåtte metoden. Neste skritt vil være å benytte mer komplekse modeller, men det medfører at optimaliserings-prosedyren blir beregningsmessig vesentlig mer krevende. I dette eksperimentet har vi brukt en modell pr. fonem-liknende enhet, såkalte monofoner. En naturlig utvidelse er å bruke kontekstavhengige modeller, såkalte trifoner, hvor vi bedre får modellert allofon-liknende enheter.
En annen naturlig utvidelse er å se på en oppgave med større vokabular. Det er også sannsynlig at det kan være en gevinst ved å kombinere optimalisering av leksikonet med oppdatering av de akustiske modellene.

8. Referanser

[1] Holter, T. og Svendsen, T., 1999, "Maximum likelihood modelling of pronunciation variation," *Speech Communication*, vol. 29 (2-4), s. 177-192.

[2] Holter, T., 1997, *Maximum likelihood modelling of pronunciation in automatic speech recognition*. Dr.ing.-avhandling, NTNU (Norges teknisk-naturvitenskapelige universitet).

[3] Kvale, K., 1996, "Norwegian numerals: A challenge to automatic speech recognition," i *Proc. ICSLP-96*, (Philadelphia, USA), s. 2028-2031.

[4] Kvale, K. og Amdal, I., 1997, "Improved automatic recognition of Norwegian natural numbers by incorporating phonetic knowledge," i *Proc. ICASSP-97,* (München, Tyskland), s. 1763-1766.

[5] Greenberg, S., 1998, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," i *Proc. Workshop*

on Modeling Pronunciation Variation for Automatic Speech Recognition, (Rolduc, Nederland), s. 47-56.
[6] Makhoul, J. og Schwartz, R., 1994, "State of the Art in Continous Speech recognition" i D. B. Roe og J. G. Wilpon, ed., "Voice Communication Between Humans and Machines" pp 165-198. National Academy Press.
[7] Amdal, I. og Ljøen, H., 1995, "TABU.0 – en norsk telefontaledatabase," Rapport nr 40/95 Telenor FoU.
[8] Höge, H. Tropf, H.S., Winsky, R., van den Heuvel, H., Haeb-Umbach, R. og Choukri, K., 1997, "European speech databases for telephone applications," i Proc. ICASSP-97 (München, Tyskalnd), s. 1771-1774
[9] http://www.phon.ucl.ac.uk/home/sampa/norweg.htm

Navn: Ingunn Amdal
Institutt: Institutt for teleteknikk
Institusjoner: NTNU/Telenor FoU
Postadresse:
Ingunn Amdal
Institutt for teleteknikk
NTNU
O.S. Bragstads plass 2B
7491 TRONDHEIM
Epost-adresse: amdal@tele.ntnu.no