FFI RAPPORT

SIGNALREPRESENTASJONER FOR AUTOMATISK TALEGJENKJENNING

GAMBORG Marius, LILLEVOLD Frode

FFI/RAPPORT-2005/01053

SIGNALREPRESENTASJONER FOR AUTOMATISK TALEGJENKJENNING

GAMBORG Marius, LILLEVOLD Frode

FFI/RAPPORT-2005/01053

FORSVARETS FORSKNINGSINSTITUTT
Norwegian Defence Research Establishment
Postboks 25, 2027 Kjeller, Norge

FORSVARETS FORSKNINGSINSTITUTT(FFI)
Norwegian Defence Research Establishment

P O BOX 25
NO-2027 KJELLER, NORWAY

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE
(when data entered)

REPORT DOCUMENTATION PAGE

1) PUBL/REPORT NUMBER	2) SECURITY CLASSIFICATION	3) NUMBER OF PAGES
FFI/RAPPORT-2005/01053	UNCLASSIFIED	31
1a) PROJECT REFERENCE	**2a) DECLASSIFICATION/DOWNGRADING SCHEDULE**	
FFI-III/876/912	-	

4) TITLE

SIGNALREPRESENTASJONER FOR AUTOMATISK TALEGJENKJENNING

Signal Representations for Automatic Speech Recognition

5) NAMES OF AUTHOR(S) IN FULL (surname first)

GAMBORG Marius, LILLEVOLD Frode

6) DISTRIBUTION STATEMENT

Approved for public release. Distribution unlimited. (Offentlig tilgjengelig)

7) INDEXING TERMS

IN ENGLISH	IN NORWEGIAN
a) Talegjenkjenning	a) Speech Recognition
b) Cepstralkoeffisienter	b) Cepstral Coefficients
c)	c)
d)	d)

THESAURUS REFERENCE: Tesaurus

8) ABSTRACT

In this report we give an overwiev of methods for front-end processing of speech signals for automatic speech recognition (ASR) that are described in the litterature.
The most common representation of speech in this context seems to be mel-frequency cepstral coeficient (MFCC) with delta- and double-delta coefficients, usually combined with cepstral mean normalization (CMN). Other representations include perceptual linear prediction (PLP) and linear prediction cepstral coefficients (LPCC).

9) DATE	AUTHORIZED BY	POSITION
2005-05-09	This page only Johnny Bardal	Director

ISBN 82-464-0936-0

FFI-B-22-1982

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE
(when data entered)

INNHOLD

Side

SIGNALREPRESENTASJONER FOR AUTOMATISK TALEGJENKJENNING

1 INNLEDNING

Tale er den mest naturlige kommunikasjonsformen mellom mennesker. I mange sammenhenger kunne det ha vært nyttig og ønskelig å bruke denne kommunikasjonsformen også overfor datamaskiner. Taleteknologi er en samlebetegnelse på teknologier som forsøker å realisere slik kommunikasjon.

Dagens taleteknologisystemer mestrer oppgaven bare delvis. Enkle oppgaver beherskes godt, men for komplekse oppgaver i vanskelige støymiljøer er teknologiens ytelse langt dårligere enn det mennesket kan prestere. Utfordringen er dermed å vurdere til hvilke oppgaver det er hensiktsmessig å bruke teknologien slik den er i dag.

FFI-prosjekt 876 SAHARA har en delaktivitet der vi holder et øye med hva som skjer på området, spesielt med tanke på mulige anvendelser for Forsvaret. Denne rapporten forsøker å gi en oversikt over den initielle signalbehandlingen som foretas i et taleteknologiprodukt. Siden teknologien fortsatt er ung og lite standardisert, har vi forsøkt å legge vekt på de mest brukte og anerkjente teknikkene som er beskrevet i litteraturen. Planen er å ta for seg den videre statistiske modelleringen i en kommende rapport.

1.1 Fagområdet taleteknologi

En liste over elementer som inngår i taleteknologi vil blant annet inneholde:

- Talegjenkjenning
- Talesyntese
- Stemmegjenkjenning
- Språkgjenkjenning
- Talekoding
- Taleforsterkning

Under talegjenkjenning forsøker man å få maskinen til å kjenne igjen det som blir sagt. Her er det et stort spenn fra gjenkjenning av enkeltord til automatisk diktering. Talegjenkjenningssystemer kan deles opp i klasser ut fra størrelsen på vokabularet de kjenner, samt om de er taleruavhengig eller trenet for stemmen og talemåten til en bestemt person.

Talesyntese er maskingenerert tale. Det brukes ofte sammen med talegjenkjenning i såkalte taledialogsystemer der man kan føre en samtale direkte med en datamaskin. I tillegg til talegjenkjenningen kreves også et delsystem som kan tolke den gjenkjente talen. Slike systemer

kan blant annet brukes til automatiske reisebestillingstjenster eller rene informasjonstjenester. De antas å ha et stort kommersielt potensiale.

Stemmegjenkjenning deles vanligvis inn i verifisering og identifisering. Ved stemmeidentifisering ønsker man å finne ut hvilken av flere kjente talere som er opphavet til en ytring. Stemmeverifisering brukes for å bekrefte at en taler er den han utgir seg for å være på bakgrunn av en ytring. Dette brukes som regel i forbindelse med aksesskontroll. Det er vanlig også å dele stemmegjenkjenning inn i tekstavhengig og tekstuavhengig gjenkjenning. I det første tilfellet leser taleren opp en allerede kjent tekst, slik at det blir lettere å utføre stemmegjenkjenningen.

Språkgjenkjenning vil si å identifisere hvilket språk som tales på bakgrunn av tale fra en ukjent taler. Slike systemer kan for eksempel tenkes brukt som en front-end for flerspråklige talegjenkjennere. De kan også benyttes i telefonsystemer til å automatisk viderekoble en oppringing til en operatør som snakker språket den som ringer benytter. Dette kan være nyttig i forbindelse med nødtelefon.

Talekoding brukes for å komprimere tale digitalt slik at den tar mindre plass. Dette er interessant i forbindelse med transmisjon av tale i telefonsystemer for å redusere båndbredden som kreves.

Det brukes ulike signalbehandlingsmetoder for å filtrere bort støy fra tale. Støyen kan ha en rekke forskjellige karakteristikker og opphav. Blant annet er det interessant å forsøke å filtrere bort bakgrunnstale fra et opptak av en samtale. Slike metoder kaller vi med en fellesbenevnelse for taleforsterkning.

1.2 Automatisk talegjenkjenning

Dagens forskning innen taleteknologi har større fokus på automatisk talegjenkjenning enn på de andre områdene. Hovedårsaken til dette er at talegjenkjenning er en av de viktigste komponentene i taledialogsystemer. I tillegg brukes de samme teknikkene innen andre deler av fagfeltet.

En vanlig oppgave for et automatisk talegjenkjenningssystem er å transkribere ytringer. De fleste systemer i bruk i dag gjør dette ved hjelp av statistiske mønstergjenkjenningsteknikker. Dette kan gjøres ved først å beregne et sett egenskaper fra signalet, og deretter sammenlikne disse med forhåndslagrede mønstre med kjent identitet. Når man gjenkjenner et mønster i signalet tilskriver man det den samme klassen som mønsterets klasse. Typiske klasser er ord og mindre talelydenheter som fonemer.

En stor utfordring er at både enkeltord og enkeltfonemer kan uttales på forskjellige måter. Uttalen varierer fra person til person, mellom dialekter og den vil til og med variere fra gang til gang for enkeltpersoner. I tillegg vil støy på grunn av det akustiske miljøet påvirke talesignalet.

For å fange opp denne uttalevariasjonen bruker man statistiske modeller av fonemene i stedet for forhåndslagrede mønstre. Samlingen av slike modeller utgjør talegjenkjenningssystemets akustiske modell. Den akustiske modellen gir en sammenheng mellom fonemer og talesignal. Modellen skal altså ivareta språkets fonetikk og fonologi. Den vanligste modelletypen i dag er skjulte markovmodeller (HMM). Man bygger en HMM for hvert fonem, og setter disse sammen til ordmodeller ved hjelp av en uttaleordbok.

I tillegg til den akustiske modellen er det vanlig å inkludere en språkmodell. Språkmodellen tar

for seg hvordan ord settes sammen til setninger. Den modellerer derfor i hovedsak språkets syntaks og semantikk, og skal øke sannsynligheten for at talegjenkjenningen fører til velformede setninger. Fordi det er vanskelig å definere en streng grammatikk for muntlig språk brukes som oftest statistisk modellering også her. Den vanligste typen språkmodell i dag er N-grammodellen.

Akustisk modell og språkmodell kan sammen brukes til å beregne sannsynligheten for at en gitt ordsekvens resulterer i et gitt talesignal. Ved talegjenkjenning søker man i de statistiske modellene etter den mest sannsynlige ordsekvensen som produserte det observerte talesignalet. Denne delen av talegjenkjenningssystemet kalles ofte for talegjenkjennerens back-end.

Talegjenkjenningssystemets front-end er det som gjøres før back-enden. Dette inkluderer registrering av talen med mikrofon og signalbehandlingen som skal til for å beregne en egnet representasjon av talesignalet. Det som er viktig for en front-end er at den gir egenskaper som gir størst mulig forskjell mellom de forskjellige fonemene. I tillegg er det viktig at den er robust for variasjoner innad for hvert fonem. Egenskapene bør også være mest mulig robuste mot støy.

1.3 Rapportens struktur

Denne rapporten fokuserer som nevnt på front-end prosessering av talesignaler for automatisk talegjenkjenning. Hensikten er å representere talesignalet med et sett egenskaper som i størst mulig grad gjenspeiler det språklige innholdet.

Kapittel 2 gir en introduksjon til talesignaler og en del egenskaper til slike signaler. Spesielt vektlegges ulike frekvensrepresentasjoner, blant annet korttidsspektre, siden de vanligste signalrepresentasjonene baserer seg på dette.

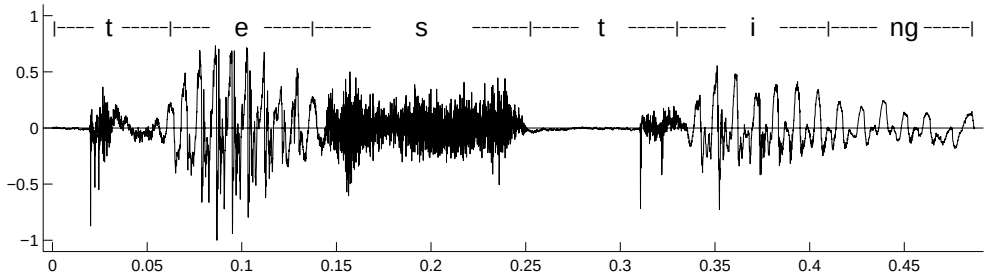
I kapittel 3 og 4 gir vi en gjennomgang av hvordan mennesker genererer og oppfatter tale. En slik gjennomgang gir innsikt om mange av talesignalets egenskaper. Dessuten er mennesket bedre enn maskiner innen de fleste aspekter som har med tale å gjøre, så det er naturlig å ta utgangspunkt i menneskets tale- og hørselsorganer når tale skal modelleres.

Lineærprediksjon er en forholdsvis gammel teknikk som kan representere talespekteret på en glatt og kompakt form (LPC). Den er basert på en forenklet modell av hvordan mennesket genererer tale, og blir presentert i kapittel 5. Andre representasjoner av tale, som er mer i bruk i dag, går ett skritt videre. De forsøker også å kompensere for hvordan menneskets hørsel fungerer. To eksempler er MFCC og PLP som blir presentert i kapittel 6.

Alle representasjonene av tale beskrevet så langt beskriver en kort ramme med data ($20 - 30$ ms). De vil dermed ikke fange opp mer saktevarierende egenskaper ved signalet. I kapittel 7 ser vi på hvordan egenskaper kan etterbehandles for å ta hensyn til slik variasjon.

2 TALESIGNALER

Et talesignal er en lydbølge som overfører verbal kommunikasjon. Taleorganene til en person skaper vibrasjoner i lufttrykket ved leppene. Vibrasjonene forplantes i form av en trykkbølge gjennom lufta, og de kan oppfattes av hørselsorganene til en mottaker.



Figur 2.1 Eksempel på et talesignal.

2.1 Grunnleggende egenskaper

Et typisk talesignal er gjengitt i figur 2.1. Langs x-aksen viser tid, mens y-aksen viser lufttrykk. Verdien 0 refererer til statisk trykk, slik at middelverdien av signalet over tid alltid blir null.

Vi ser at signalet veksler mellom nestenperiodiske deler med stor amplitude og energi, og mer støylignende deler med mindre energi. Periodisiteten kommer fra stemmebåndene som vibrerer med en grunnfrekvens f_0 for stemte lyder. Ved ustemte lyder er det ingen vibrasjon, og luften passerer stemmebåndene som en turbulent strømning og gir et signal som ligner hvit støy. Grunnfrekvensen f_0 ligger i området $60 - 400$Hz. Mer informasjon om generering av talesignaler fins i kapittel 3.

Nesten all informasjonen i tale ligger i frekvensområdet under 8 kHz, med klar hovedtyngde i den laveste halvdelen av båndet. Ved opptak av tale med høy kvalitet er det vanlig å benytte en samplingsrate på minst 16 kHz og bruke minst 16 bit per sampel. For telefonkvalitet benyttes oftest 8 kHz og 8 bit. Det fins imidlertid mange ulike dataformater for tale. Mange av dem bruker ulike former for komprimering.

Et talesignal er beheftet med ulike typer støy. De to hovedgruppene er additiv støy og foldningsstøy. Additiv støy er bakgrunnslyder som kommer til øret samtidig med talen. Ofte modellerer man denne støyen som hvit, men tilnærmelsen er vanligvis ikke spesielt god. Foldningsstøy er forvrenginger av talesignalet som skjer under overføringen. Det kan for eksempel være ekkoeffekter eller filtrering i forbindelse med telefonoverføring. I de fleste modeller antar man at kanalen som overfører signalet kan beskrives som et lineært tidsinvariant filter.

En annen utfordring i forbindelse med prosessering av talesignaler er den komplekse sammenhengen mellom den fysiske beskrivelsen av signalet og hvordan lyden blir oppfattet av mennesker. Ofte kan tilsynelatende meget forskjellige signaler høres like ut. Noe av utfordringen er derfor å representere talesignaler på en måte som kan hjelpe til med å beskrive hvordan mennesker oppfatter dem.

2.2 Korttidsspekter

Det er vanskelig å trekke relevant informasjon ut fra tidsrepresentasjonen av et talesignal $s(t)$. Flere ting taler for en frekvensrepresentasjon av signalet. Kilde-filtermodellen for taleproduksjon (avsnitt 3.2) får en enklere beskrivelse i frekvensplanet og øret utfører i prinsippet en

frekvensanalyse av det innkomne signalet. Det er gjort undersøkelser som tyder på at øret i enkelte tilfeller kan oppfatte forvrengning av fasespekteret til signalet. Likevel konkluderer disse med at slik forvrengning generelt ikke vil oppfattes (12). Mange nyttige beskrivelser av talesignaler tar derfor utgangspunkt i effektspekteret.

Frekvensene i et talesignal varierer over tid. Over korte tidsrom er signalet likevel forholdsvis stasjonært eller periodisk. Det er derfor naturlig å beskrive signalet ved hjelp av en vindusbasert fouriertransform

$$S(f,t) = \int_{-\infty}^{\infty} s(\tau)w(\tau - t)e^{-2\pi j f \tau}\,\mathrm{d}\tau. \tag{2.1}$$

En slik representasjon bruker først et vindu $w(t)$ til å avgrense signalet i tid, og tar deretter fouriertransformen av det avgrensede signalet. Siden faseinformasjonen som nevnt ikke har stor betydning for taleoppfattelsen brukes stort sett korttidsspekteret gitt ved

$$P_S(f,t) = |S(f,t)|^2 \tag{2.2}$$

I følge Heisenbergs usikkerhetsrelasjon (16) kan man ikke velge et vindu $w(t)$ som gir god oppløsning i både tid og frekvens samtidig. I vårt tilfelle ønsker vi en god frekvensoppløsning, men må samtidig ta hensyn til den underliggende antagelsen om at talesignalet er tilnærmet periodisk eller stasjonært innenfor vinduet. Denne antagelsen blir dårligere for lengre vinduer. Som et kompromiss brukes ofte et Hamming- eller Hanningvindu (10) med lengde på 20-30 ms. Vi sier mer om betydningen av vinduslengden i neste avsnitt.

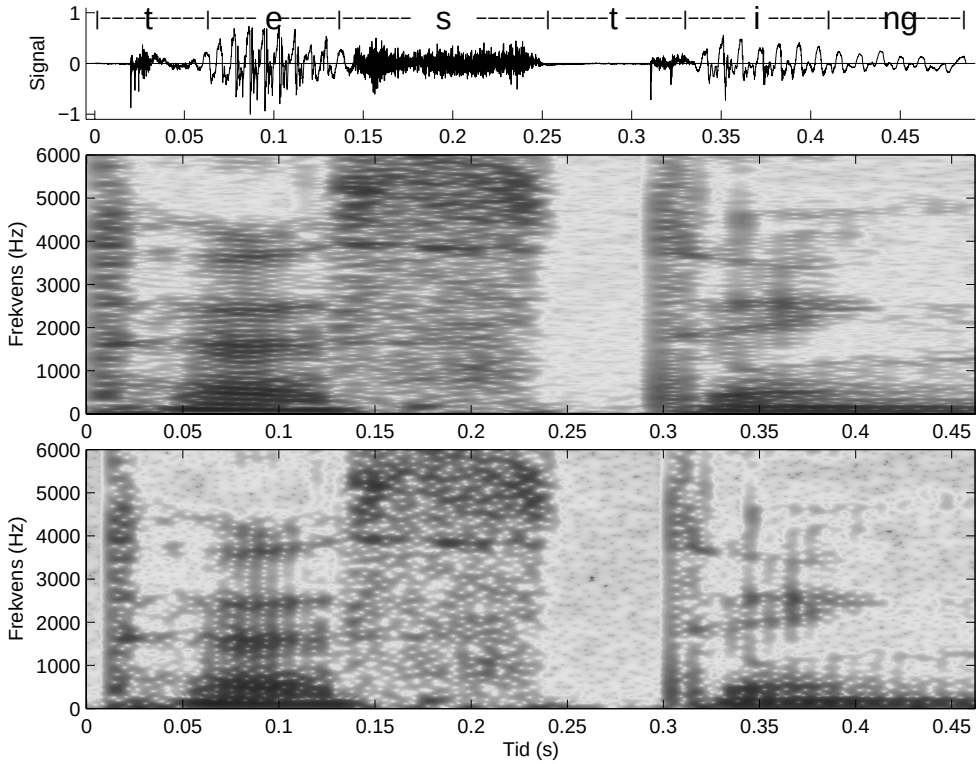
2.3 Spektrogram

Et spektrogram er en grafisk framstilling av hvordan korttidsspekteret til et signal endrer seg over tid. Det er et viktig redskap innen taleprosessering. Spektrogrammet har en tidsakse og en frekvensakse, og verdien av korttidsspekteret $P_S(f,t)$ er plottet med en farge- eller gråtoneskala. Et eksempel er gitt i figur 2.2.

Det man hovedsakelig kan lese ut av et talespektrogram er hvordan formantfrekvensene i signalet endrer seg over tid. Disse gjenkjennes som frekvensbånd med mye energi over en viss tidsperiode. For et trenet øye er informasjonen i spektrogrammet nok til å "lese" signalet.

Spektrogrammer kan klassifiseres som smalbåndet eller bredbåndet alt etter hvor bredt vinduet $W(f)$, som blir brukt under beregning av spekteret, er i forhold til grunnfrekvensen f_0 i signalet. I et bredbåndsspektrogram som det nederst i figur 2.2, bruker vi et vindu som i tid er smalere enn en periode til grunnfrekvensen. Resultatet blir at grunnfrekvensen vises i tidsoppløsningen til spektrogrammet som loddrette streker separert med en periode. Et smalbåndsspektrogram er vist øverst i figur 2.2. Her er frekvensoppløsningen så god at vinduet følger de individuelle resonanstoppene fra grunnfrekvensen. Resultatet blir vannrette linjer separert med en frekvensavstand f_0.

Hvis man tilpasser vinduet til grunnfrekvensen får man et såkalt pitsj-synkront spektrogram. Da skal resonansene med frekvens f_0 være skjult både i tid og frekvens. I praksis er det imidlertid problematisk å estimere og følge en grunnfrekvens som kan variere over tid. Vanligvis leses formantene fra et bredbåndsspektrogram. Der kan ikke formantene forveksles med resonanslinjer stammende fra f_0.



Figur 2.2 Talesignal, smalbåndsspektrogram og bredbåndsspektrogram.

3 GENERERING AV TALE

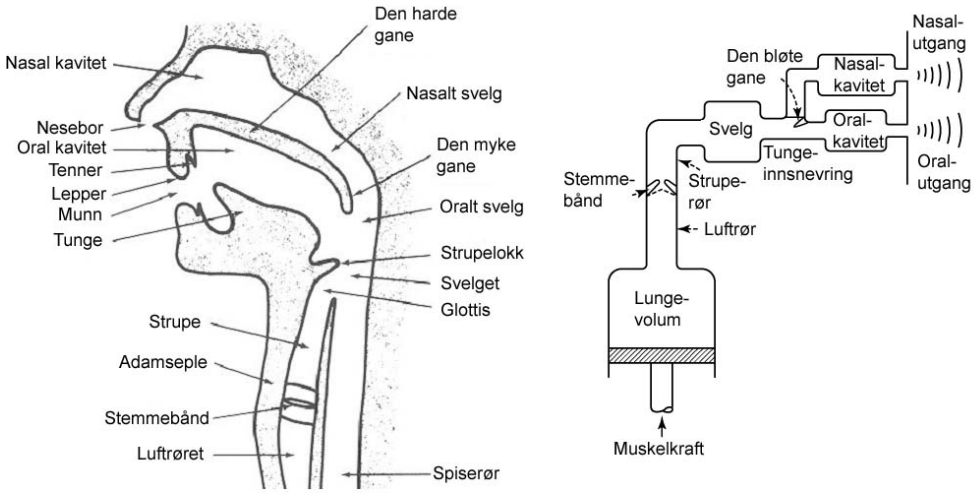
3.1 Taleorganene

Tale genereres ved hjelp av et system av taleorganer. Disse organene er presentert i figur 3.1. Det er vanlig å dele dette systemet inn i to deler, det underglottale- og det overglottale systemet. Det overglottale systemet kalles ofte for vokaltrakten eller ansatsrøret. Ansatsrøret består igjen av halsen, nasalkaviteten og oralkaviteten.

Det underglottale systemet består av lunger, luftrør og stemmebånd. Dette systemet fungerer som energikilden til taleproduksjonssystemet. Lungene presser luft opp igjennom stemmebåndet og videre til ansatsrøret. Denne luftstrømmen refereres ofte til som systemets eksitasjon eller grunnsignal. Enkelte talelyder kan også produseres ved hjelp av andre kilder, men det er ikke tilfellet for noen av de vanlige talelydene i norsk og engelsk. Eksempler på dette er klikke- og smattelyder man av og til bruker når man skal uttrykke misnøye.

Stemmebåndene bestemmer formen til grunnsignalet. Dersom stemmebåndene ikke strammes vil en kontinuerlig luftstrøm slippes igjennom til ansatsrøret. På grunn av innsnevring ved stemmebåndene vil luftstrømmen bli turbulent. Dette resulterer i et støylignende grunnsignal med et flatt frekvensspekter. Grunnsignalet kan i dette tilfellet kan sees på som hvit støy. Tale som genereres på denne måten kalles for ustemt tale.

Dersom stemmebåndene strammes vil det sperre lufttilførselen til ansatsrøret. Når lungene presser på vil dette føre til økt trykk ved stemmebåndene og de vil etterhvert presses fra hverandre. Luft



Figur 3.1 De viktigste taleorganene hos mennesket. Venstre halvdel av figuren viser en skisse av hele det overglottale- og deler av det underglottale systemet. Høyre halvdel viser en prinsippskisse av hele systemet.

slippes da igjennom, trykket reduseres og stemmebåndene lukkes igjen. Dette vil gjentas mange ganger hvert sekund slik at stemmebåndene vibrerer. Når talen genereres på denne måten har vi stemt tale. Det resulterende grunnsignalet er en nestenperiodisk pulset luftstrøm. En periode av dette signalet kalles for en glottalpuls. Noen skisser av stemte grunnsignal er vist i figur 3.2.

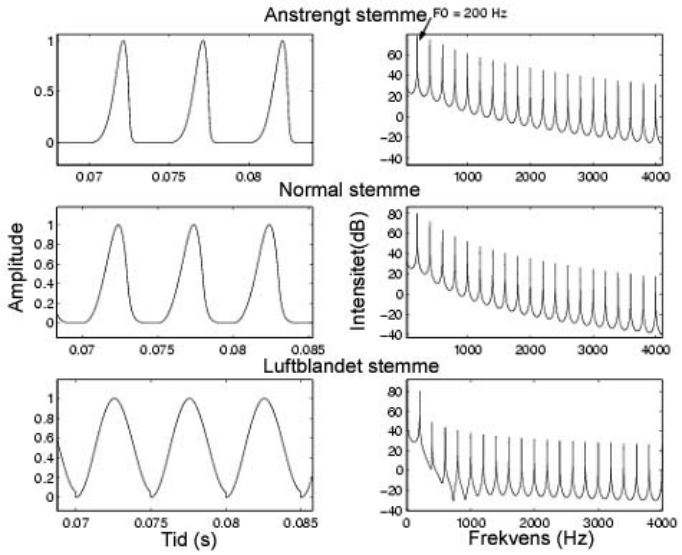
Perioden til et stemt grunnsignal bestemmer grunnfrekvensen f_0. Grunnfrekvensen kan være så lav som 60 Hz for menn og høyere enn 300 Hz for kvinner og barn (10). Frekvensen bestemmes av stemmebåndets størrelse og elastisitet i tillegg til styrken på lufttrykket som påtrykkes fra lungene. Disse faktorene kan til en viss grad påvirkes av taleren slik at grunnfrekvensen kan varieres noe.

Ved stemt tale har grunnfrekvensen stor betydning for frekvensinnholdet i talesignalet. Spekteret inneholder hovedsakelig spektrallinjer med frekvenser som er et helt multiplum av grunnfrekvensen. Spektrallinjene ligger tettere desto lavere grunnfrekvensen er.

Grunnsignalet propagerer videre igjennom ansatsrøret og ut av munnen. Inngangen til nasalkaviteten kan åpnes og lukkes ved hjelp av den bløte ganen. Dersom den åpnes vil noe av luftstrømmen gå igjennom nasaltrakten og ut av neseborene. Talesignalet er de resulterende trykkbølgene som kommer ut av munn og nese.

Ansatsrøret er formet som et rør med en 90 grader bøy cirka midt på. Selve formen kan varieres over tid, og dette gjøres ved hjelp av artikulatorene. De viktigste artikulatorene er lepper, tenner, tunge, harde gane og bløt gane. Disse vil påvirke lydsignalet på veien gjennom ansatsrøret. Særlig der det er store endringer i ansatsrørets tykkelse vil noe av lyden reflekteres. Refleksjoner vil kunne føre til at enkelte frekvenser i lydsignalet forsterkes og andre svekkes. Ansatsrøret vil med andre ord påvirke frekvensinnholdet i talesignalet. Som regel vil noen få frekvenser forsterkes mye. Disse kalles resonanser, resonansfrekvenser eller formanter. Formantene benevnes vanligvis f_1, f_2 og så videre, etter stigende frekvens.

Når den bløte gane er åpen slik at luft kan strømme gjennom nasalkaviteten vil det oppstå såkalte



Figur 3.2 Noen eksempler på stemt grunnsignal med tilhørende plott av signalets frekvensinnhold.

nasale lyder. I tillegg til å ha formanter er disse lydene preget av at enkelte frekvenser vil være kraftig dempet. Denne dempningen er et resultat av destruktiv interferens mellom lyd i oralkaviteten og i nasalkaviteten. Frekvensene som dempes bort kalles ofte antiresonanser.

3.2 Modell av talegenerering

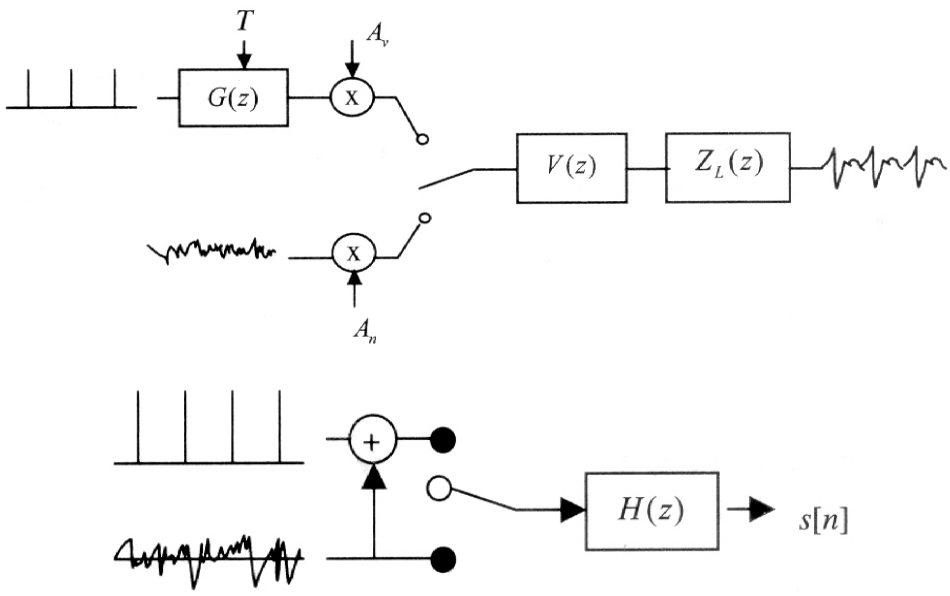
Når tale skal modelleres ønsker man at modellen skal være enklest mulig. Samtidig må de deler av talesignalet som er viktig for oppfattelsen av tale ivaretas av modellen. Videre er det ønskelig at modellen tar utgangspunkt i den fysiske talegenereringsprosessen, selv om dette strengt tatt ikke er nødvendig.

Den vanligste modellen deler talegenereringsprosessen i en kildedel som skaper grunnsignalet og en filterdel som farger signalet. En illustrasjon av modellen er gitt i øvre del av figur 3.3.

Som nevnt deler man inn talen i stemt og ustemt tale. På grunn av dette brukes to forskjellige kilder i modellen. Det ustemte grunnsignalet er hvit støy. Det stemte signalet genereres fra et impulstog med frekvens lik grunnfrekvensen som foldes med glottalpulsen. Filteret som beskriver glottalpulsen er benevnt $G(z)$ i figur 3.3. $V(z)$ er ansatsrørets filterfunksjon og $Z_L(z)$ er impedansen til leppene.

Modellen forenkles ofte ved å bruke et enkelt filter $H(z)$ som det eneste filteret i modellen. Dette filteret er da kombinasjonen av $G(z)$, $V(z)$ og $Z_L(z)$ for stemt tale, men bare $V(z)$ og $Z_L(z)$ for ustemt tale. For å ta hensyn til at noen talelyder inneholder en kombinasjon av stemt og ustemt lyd kan de to grunnsignalene kombineres additivt. Resultatet blir da modellen nederst i figur 3.3.

Filteret modelleres som et lineært filter som varierer sakte over tid. For kontinuerlige talelyder vil filteret være tilnærmet stasjonært over tidsperioder av størrelsesorden 20 ms. Dette er bare en tilnærming, og fenomener som koartikulasjon vil føre til at antakelsen ikke er helt rett.



Figur 3.3 Illustrasjon av kilde-filtermodellen, fra (10).

Det er vanlig å modellere filteret som et minimum-fase all-pole system. Dette er en forholdsvis enkel modell som fanger opp mange av egenskapene til det virkelige filteret. Årsakene til at et slikt filter velges som modell blir forklart ytterligere i avsnitt 5.1.

Fordi man modellerer filteret som et all-pole filter får det problemer med å modellere nasale lyder. Nasale lyder har som nevnt tilnærmet utslukking av enkelte frekvenser, slik at filteret $H(z)$ burde inneholde nullpunkter. Modellen kan likevel være en god tilnærmelse dersom nok poler brukes da et nullpunkt kan uttrykkes ved et uendelig antall poler.

4 OPPFATTELSE AV TALE

Mennesket har en meget god evne til å oppfatte tale, selv når talesignalet har et signal-støy-forhold (SNR) under 0 dB. Mange automatiske talegjenkjenningssystemer forsøker å etterlikne det menneskelige hørselssystem for å prøve å oppnå noe av den samme støyrobustheten.

Det er to måter å få innblikk i hørselssystemet på. Den ene måten er å se på fysiologien til hørselsorganene. Den andre er å studere hørselen ved hjelp av psykoakustiske eksperimenter. Det vil si å undersøke hvordan mennesket oppfatter forskjellige lyder.

4.1 Ørets fysiologi

Det menneskelige hørselssystemet fanger opp trykkbølger, frekvensanalyserer disse og sender nerveimpulser via hørselsnerven til prosessering i hjernen. Det er vanlig å dele inn hørselssystemet i to hoveddeler, de indre- og ytre hørselsorganer.

De indre hørselsorganer omfatter hørselsnervesystemet og hjernes oppfattelse av lydsignalet. Man

Fysisk størrelse	Perseptuell kvalitet
Lydintensitet	Lydstyrke
Grunnfrekvens	Pitsj (stemmeleie)
Spektral omhylningskurve	Klangfarge
Faseforskjell (stereo)	Lydkildeposisjon

Tabell 4.1 Fysiske størrelser og nært beslektede perseptuelle kvaliteter.

har lite kunnskap om hvordan hjernen behandler signalene. Derfor legger de fleste automatiske talegjenkjenningssystemer vekt på å etterlikne hvordan de ytre hørselsorganene fungerer, med andre ord hvordan øret overfører lydsignalene til nerveimpulser.

De ytre hørselsorganene kan deles inn i det indre øret, mellomøret og det ytre øret. Det ytre øret overfører lydbølgene til trommehinnen. Mellomøret fungerer som en overgang fra trykkbølger til mekaniske vibrasjoner, og overfører disse til det indre øret. Disse to delene av hørselssystemet har i hovedsak som oppgave å forsterke lydsignalet, og å utføre impedanstilpasning mellom forplantningsmediene i mellom- og indre øre.

Det indre øret består av sneglegangen og de indre hårcellene, som sender signaler til hørselsnerven. Sneglegangen, også kalt cochlea, er et spiralformet væskefylt rør som er delt i lengderetningen av basilærmembranen. Vibrasjoner fra lydbølgene forplantes via det ovale vinduet til væsken som fyller sneglegangen. Bølgene i væsken vil igjen føre til utslag på basilærmembranen. Posisjonen til disse utslagene avhenger av lydens frekvens. Membranen er følsom for høye frekvenser nærmest det ovale vindu og for lave frekvenser lengre vekk. Avstanden fra enden av membranen til utslaget er omtrent proporsjonalt med logaritmen til frekvensen til lyden som skaper utslaget. Frekvensoppløsningen er høyere for lave frekvenser enn for høye.

En god modell av det indre øret er derfor en båndpassfilterbank, hvor ulike filtre tilsvarer ulike posisjoner langs basilærmembranen. De enkelte filtrene bør ha båndbredde som er proporsjonal med frekvens.

De indre hårcellene reagerer på vibrasjonene på basilærmembranen og sender ut nerveimpulser. Mengden nevral aktivitet øker med økende amplitude til utslaget på basilærmembranen. Dette fører til at den nevrale aktiviteten er omtrentlig proporsjonal med logaritmen til lydintensiteten. Oppfattelse av lydstyrke er derfor logaritmisk i forhold til lydintensitet. I tillegg er frekvensfølsomheten dårligere for høye frekvenser.

4.2 Psykoakustiske fenomener

Psykoakustikk dreier seg om forholdet mellom lydsignalets fysiske egenskaper og hvordan det oppfattes av mennesker. For å tydeliggjøre dette skiller man vanligvis mellom lydsignalets fysiske og perseptuelle egenskaper. Tabell 4.1 viser de viktigste fysiske størrelser og de tilhørende perseptuelle kvaliteter. Det er stor grad av korrelasjon mellom disse egenskapene, men det er absolutt ikke et en-til-en forhold mellom dem. En gitt perseptuell kvalitet kan påvirkes av flere fysiske størrelser gjennom komplekse sammenhenger.

En lyd med høy effekt vil generelt høres sterkere ut. Likevel vil lyder med forskjellig frekvens men samme effekt ikke nødvendigvis høres ut som om de er like sterke. Ørets følsomhet varierer altså med frekvensen.

På samme måte er det med grunnfrekvens og pitsj. Dersom en lyd har konstant grunnfrekvens mens intensiteten endres vil man som regel oppfatte en viss variasjon i pitsj.

Klangfarge er definert ut ifra lydstyrke og pitsj. Begrepet omfatter alle kvaliteter ved en lyd som gjør at den kan skilles fra en annen lyd med samme pitsj og lydstyrke.

Det har blitt gjort en rekke studier av psykoakustiske fenomen. De fleste av disse er basert på hvordan det indre øret frekvensfiltrerer lyden, og hvordan lydoppfattelsen er i forhold til frekvens og lydintensitet.

4.3 Frekvensfølsomhet

Oppbygningen av det indre øret tyder på at ørets oppfattelse av toner ikke er lineær i forhold til lydens frekvens. Man har gjort forsøk for å finne en frekvensskala som modellerer hørselssystemets naturlige frekvensfølsomhet. Ved hjelp av ulike forsøk er det funnet flere slike skalaer.

En av disse kalles mel-skalaen (10). Denne er lineær for frekvenser under 1 kHz og logaritmisk over. Mel-skalaen er basert på forsøk der man har spilt av en enkeltone og bedt forsøkspersoner om å justere tonen til den høres halvparten så høy ut. Skalaen er definert slik at man har 1000 mel ved 1 kHz. En tilnærming til skalaen er gitt ved

$$M(f) = 2295 \log(1 + f/700) \tag{4.1}$$

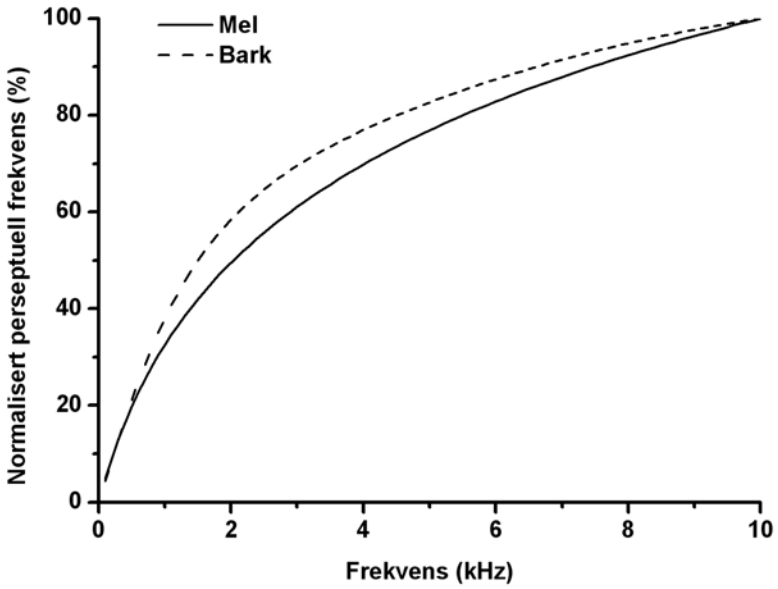
I likningen er M mel-skala verdi, f frekvens og logaritmen er med grunntall 10.

Mel-skalaen er mye brukt i moderne talegjenkjenningssystemer. En av de vanligste egenskapene brukt innen talegjenkjenning, MFCC, bruker denne skalaen. MFCC er beskrevet nærmere i 6.2.

En annen perseptuell frekvensskala er bark-skalaen (10). Bark-skalaen er basert på ideen om kritiske bånd. Fletcher (3) har påpekt at det eksisterer kritiske bånd i frekvensresponsen til basilærmembranen. Disse båndene representerer hørselssystemets evne til å skille mellom samtidige toner med lik effekt. Hvis frekvensene ligger for nær hverandre, altså innen en kritisk båndbredde, vil de ikke høres ut som separate toner. Årsaken til dette ligger i at lydene påvirker nærliggende områder på basilærmembranen, slik at utslagene på den overlapper. Undersøkelser har vist at lydene ikke vil kunne skilles fra hverandre dersom områdene de påvirker på basilærmembranen ligger innen ca 1,5 mm av hverandre. Det viser seg at basilærmembranens respons fungerer omtrent som en filterbank med 24 overlappende filtre med båndbredde lik den kritiske båndbredden. Bark-frekvens er derfor definert på en skala fra 1 til 24 slik at den kritiske båndbredden alltid er lik 1 bark. Bark-frekvens kan tilnærmes som

$$b(f) = 13 \arctan(0,00076\, f) + 3,5 \arctan((f/7500)^2) \tag{4.2}$$

Et plott av mel- og bark-skalaene er vist i figur 4.1. Skalaene er såpass like at det spiller veldig liten rolle hvilken av skalaene som benyttes i en automatisk talegjenkjenner.



Figur 4.1 Mel- og bark-skalaene plottet som funksjon av lydfrekvens.

4.4 Lydstyrke

Lydstyrke er en psykoakustisk størrelse som sier noe om hvor høy en lyd oppfattes. Sammenhengen mellom lydstyrke og lydeffekt finnes ved å undersøke hvor mye effekten må økes for at en forsøksperson skal oppfatte det som en dobling av lydstyrken. Det er vanlig å utføre slike målinger i stille rom med en tone på 1 kHz.

Det viser seg at for lyder over 40 dB så dobles lydstyrken ved en økning på cirka 10 dB. Det betyr at lydstyrke kan modelleres som proporsjonalt med kubikkroten av lydeffekten.

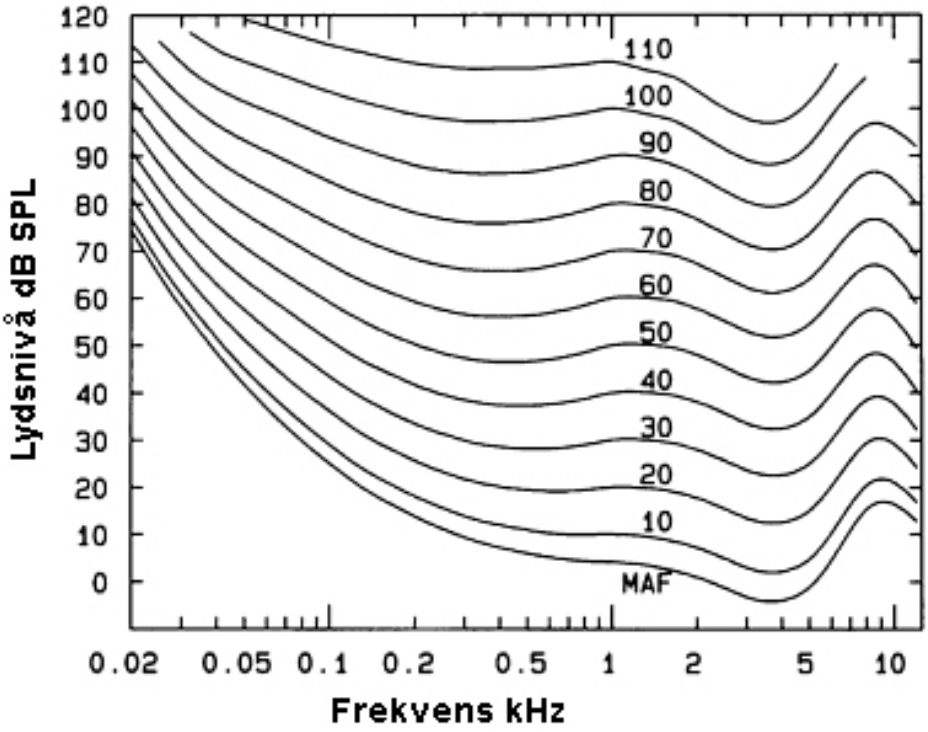
For lyder under 40 dB vil forholdet endres slik at det skal en mindre endring i intensitet til for å doble lydstyrken.

Oppfattelsen av lydstyrke avhenger som nevnt av frekvens, og er beskrevet i standarden ISO 226: Normal equal-loudness-level contours. Standarden angir såkalte “equal loudness curves”, som viser hvor høy lydintensitet som kreves ved forskjellige frekvenser for at lydene skal oppfattes som like høye. Et eksempel på slike kurver er vist i figur 4.2.

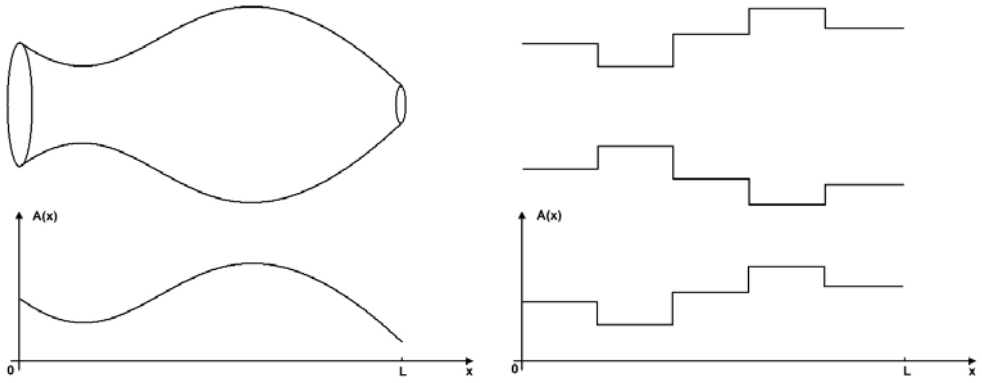
5 LINEÆRPREDIKSJON

5.1 Rørkjedemodell av ansatsrøret

Som nevnt i seksjon 3 kan det menneskelige talegenereringssystemet sees på som et todelt system, bestående av signalkilde og filter. Endringer i tverrsnittsarealet til ansatsrøret påvirker motstanden mot lydbølgene. Endring i motstand vil føre til refleksjoner av lydenergien. Ved



Figur 4.2 Equal loudness kurver. Hver kurve viser hvilket lydnivånivå i dB som kreves for at lydstyrken ved en gitt frekvens skal tilsvare lydstyrken ved et gitt lydnivå for en tone med frekvens 1 kHz.



Figur 5.1 Illustrasjon av rørkjedemodellen for ansatsrøret. Venstre del viser en kontinuerlig arealfunksjon, mens høyre viser rørkjedemodellen med fem rør.

utgangen av ansatsrøret vil noen frekvenskomponenter forsterkes og andre svekkes på grunn av interferens mellom de ulike reflekterte komponentene av lydsignalet.

Ansatsrørets variasjon i tykkelse som funksjon av posisjon langs røret kalles ansatsrørets arealfunksjon. For å modellere ansatsrøret på en enkel måte kan man anta at ansatsrøret er en sammensetning av N antall like lange tapsfrie rør med forskjellig tykkelse. Verdien på arealfunksjonen vil da kun endre seg i overgangen mellom rør, og det er kun her refleksjoner vil finne sted. I virkeligheten vil selvsagt arealfunksjonen være kontinuerlig. At rørene er tapsfrie vil si at man ser bort fra tap av lydenergi på grunn av friksjon, varmeledning samt vibrasjoner som forplantes til veggene i ansatsrøret. Dersom man bruker et tilstrekkelig stort antall rør vil dette være en god modell for hvordan ansatsrøret filtrerer lydsignalet. En illustrasjon av modellen er vist i figur 5.1.

Modellen forenkles ytterligere ved å anta at bølgeutbredelsen kun skjer langs lengderetningen til ansatsrøret. Dette er en god tilnærmelse for bølgelengder som er store i forhold til ansatsrørets tykkelse. I praksis kan man bruke tilnærmelsen for frekvenser lavere enn 4 kHz.

Ved hver røroverganger vil andelen lydenergi som reflekteres være gitt av forskjellen på arealfunksjonen til de to rørene. Refleksjonskoeffisientene er gitt ved

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{5.1}$$

Dette gir refleksjonskoeffisienter mellom -1 og 1, hvor fortegnet avgjør refleksjonens fase. Filterfunksjonen for et ansatsrør satt sammen av N rør vil da kunne utledes ved å kreve kontinuerlig trykk og volumstrømning ved alle røroverganger.

$$V(z) = \frac{U_L(z)}{U_G(z)} = \frac{0.5z^{-N/2}(1+r_G)\prod_{k=1}^{N}(1+r_k)}{[1\ -r_G]\left(\prod_{k=1}^{N}\begin{bmatrix} 1 & -r_k \\ -r_k z^{-1} & z^{-1} \end{bmatrix}\right)\begin{bmatrix} 1 \\ 0 \end{bmatrix}} \tag{5.2}$$

Her representerer k rørnummeret, der rørene nummereres i stigende rekkefølge fra stemmebåndet mot leppene, r_G er refleksjonskoeffisienten ved stemmebåndet, mens r_k er koeffisienten for

refleksjon mellom rør nummer k og nummer $k+1$. Likningen gir en filterfunksjon med $N/2$ nullpunkter i origo og $N/2$ komplekskonjugerte poler. Nullpunktene i origo gir bare en forsinkelse av signalet. Hver pol representerer en resonans i systemet, altså en formant.

Eksperimenter viser at ansatsrøret vanligvis har maksimalt en formant per kHz båndbredde (2). For praktisk bruk av modellen bør derfor antall poler p settes til minst $F_s + 2$, der F_s er samplingsraten i kHz. Det vil si at man trenger minst $F_s/2 + 1$ rør i rørmodellen.

Fordi nullpunktene i origo kun gir en forsinkelse er det vanlig å se bort ifra dem. Dermed kan man se på filterfunksjonen som et filter som kun inneholder poler.

Effekten av termisk energitap i ansatsrøret vil være en liten økning i resonansfrekvensene. Forplantning av vibrasjoner til ansatsrørets vegger vil ha den motsatte effekten. Den samlede effekten vil i praksis gjøre resonansene noe bredere.

For nasale lyder vil nasaltrakten være akustisk koblet til vokaltrakten. Systemet vil derfor være et rør som forgreines i to mot enden. Dette systemet er vesentlig forskjellig fra rørkjedemodellen. Ved en del frekvenser vil refleksjoner fra nasal- og oraltraktene oppheve hverandre, slik at man også får anti-resonanser i filterfunksjonen. Rørkjedemodellen kan fungere også for nasale lyder dersom man bruker mange nok rør.

5.2 Lineærprediksjonskoeffisienter (LPC)

Lineærprediksjon baserer seg på ideen om at et nytt sampel av signalet kan tilnærmes ved en vektet sum av tidligere sampler (17). Grunnlaget for bruk av metoden på talesignalet er kilde-filtermodellen beskrevet i seksjon 3.2. I forrige avsnitt ble det antydet at et filter $H(z)$ med bare poler er en god representasjon av ansatsrøret i modellen. Et filter med p poler kan skrives på den generelle formen

$$H(z) = \frac{S(z)}{E(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5.3}$$

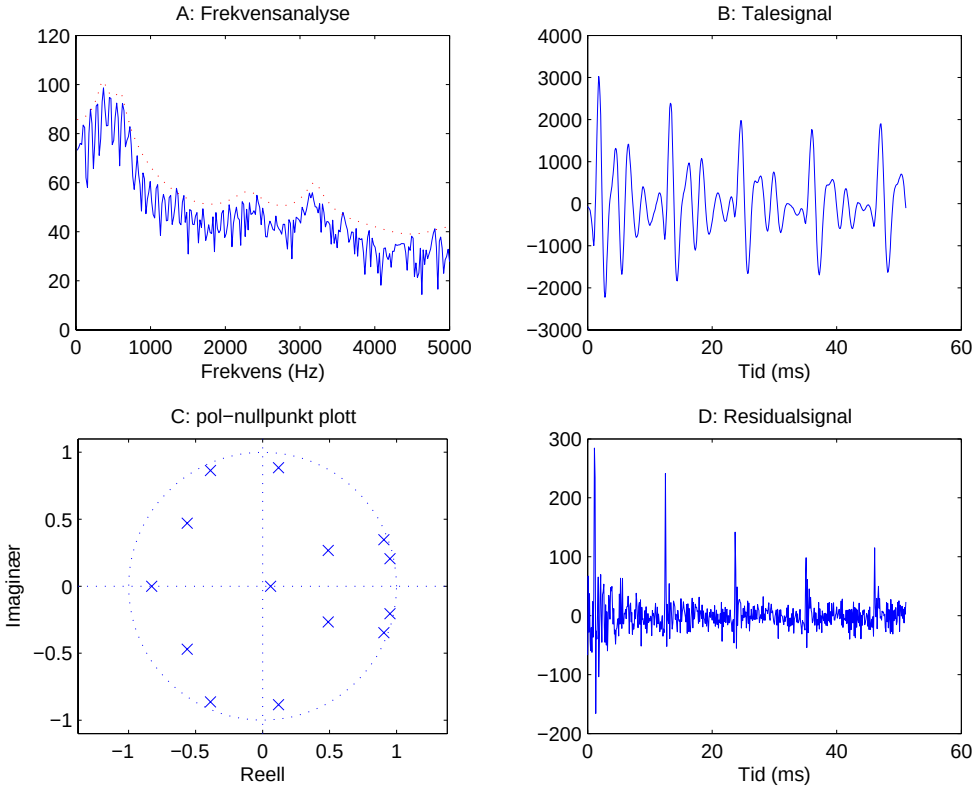
I likningen er $E(z)$ z-transformen av eksitasjonssignalet og $S(z)$ er z-transformen av talesignalet. Ved å ta invers z-transform får man

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n] \tag{5.4}$$

For en modell der filteret har p poler og ingen nullpunkter er altså signalet ved sampel n en vektet sum av tidligere sampler pluss eksitasjonssignalet. Hvis man bruker lineærprediksjon av orden p ønsker man å estimere signalet ved sampel n fra de p foregående sampler. For det predikerte signalet $\tilde{s}[n]$ får vi dermed

$$\tilde{s}[n] = \sum_{k=1}^{p} \alpha_k s[n-k] \tag{5.5}$$

Differansen mellom talesignalet og lineærpredikert signal kalles residualet. Dersom talesignalet følger kilde-filtermodellen med $\alpha_k = a_k$, så er residualet lik eksitasjonssignalet $e[n]$.



Figur 5.2 *A viser et 14-pols LPC-spekter (stiplet linje) for talesegmentet i B plottet sammen med FFT-spekteret for samme talesegment. C viser polenes plassering i enhetssirkelen, og D viser residualsignalet.*

Ved lineærprediksjon ønsker man å bestemme de LPC-koeffisientene α_k som best modellerer talesignalet. Fordi talesignalet varierer over tid vil det ikke gi noen mening å estimere LPC-koeffisientene for lange perioder. I praksis estimerer man LPC-koeffisientene ved å minimere kvadratsummen av feilen E_n over korte segmenter av talesignalet. Dette er vist i følgende likning, der summen tas over m sampler fra talesignalet.

$$E_n = \sum_m e[n]^2 = \sum_m \left(s[n] - \sum_{k=1}^{p} \alpha_k s[n-k] \right)^2 \tag{5.6}$$

Eksitasjonssignalet ved stemt tale er et pulstog med periode bestemt av grunnfrekvensen. Dette signalet er lik null mesteparten av tiden, så bruk av minste kvadraters metode vil fungere godt. En annen grunn til å bruke denne metoden er at den leder frem til et sett med lineære likninger på en form som kan løses effektivt.

Resultatet av LPC-analysen er et allpolfilter som er tilpasset korttidsspekteret til talesignalet for den gitte analyserammen. Et eksempel på dette er vist i figur 5.2 for vokalen “a”.

Dersom antakelsen om allpolfilter stemmer, vil residualsignalet være lik eksitasjonssignalet. Det vil si at det for stemt tale ligner et impulstog med periode gitt av grunnfrekvensen. I figur 5.2 er talesegmentet tatt fra stemt tale, og residualsignalet likner et impulstog. For ustemt tale vil residualsignalet likne hvit støy.

5.3 Lineærprediksjon i talegjenkjenning

Lineærprediksjon brukes i dag mye inne talekoding, men kan også brukes som basis for egenskapsuttrekning innen talegjenkjenning. Metoden baserer seg på lineærprediksjon av orden p på et talesegment av lengde N. I de fleste tilfeller brukes orden p i området 8 til 14, med en lengde på talesegmentet på 20 til 30 ms. Antall sampler i talesegmentet vil dermed avhenge av samplingsraten.

Metoden gir en god representasjon av den spektrale omhylningskurven til talesignalet dersom man bruker en passende orden p. Dersom p er for lav vil man ikke modellere alle de spektrale toppene godt nok. For høy p vil føre til at modellen tar med seg informasjon om tilfeldig variasjoner i signalet som kan skyldes støy.

For bruk innen automatisk talegjenkjenning fins det flere representasjoner av talesignalet som tar utgangspunkt i LPC. Disse brukes imidlertid i liten grad siden de er følsomme for støy på talesignalet (11). Den eneste metoden som til en viss grad brukes i dag kalles LPCC (Linear Prediction Cepstral Coefficients). Metoden beregner cepstralkoeffisienter (se kapittel 6) fra LPC ved hjelp av en rekursiv algoritme. Omregningen til cepstralkoeffisienter gjøres for å dekorrelere egenskapsvektoren.

6 CEPSTRALKOEFFISIENTER

6.1 Cepstrum

Vi har allerede sett at metoder for separasjon av kilde og filter er viktig innen taleprosessering. Kilden representerer i denne sammenhengen det signalet som blir generert ved stemmebåndene, mens filteret representerer den omformingen av signalet som skjer ved hjelp av taleorganene. De mest brukte metodene for slik separasjon baserer seg på cepstralprosessering av signalet.

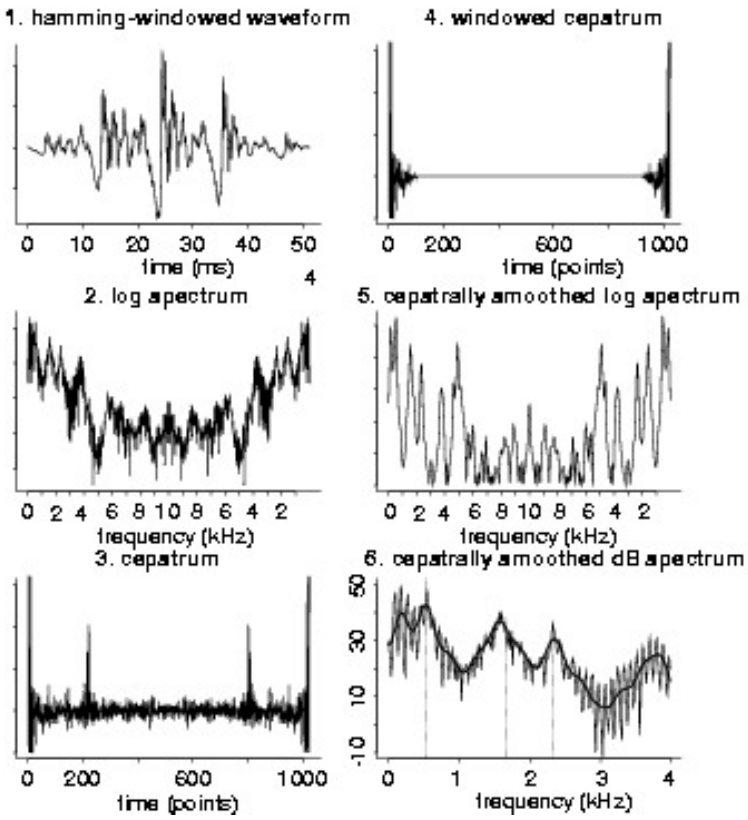
Selve ordet “cepstrum” stammer fra (1) og er en omskrivning av “spectrum”. Bakgrunnen for navnet er at man bruker signalbehandlingsmetoder for tidssignaler på et frekvenssignal, og dermed snur opp-ned på begrepene. Som vi skal se kan cepstralanalyse sees på som filtrering av log-spekteret til et signal.

Det reelle cepstrumet til et signal $s(t)$ finner man ved å fouriertransformere signalet, ta logaritmen til absoluttverdien, og til slutt ta invers fouriertransform.

$$s(\tau) = \mathcal{F}^{-1} \log |S(f)| \tag{6.1}$$

På grunn av absoluttverdien er ikke denne transformen inverterbar, noe det komplekse cepstrumet som bruker kompleks logaritme er. Innen taleanalyse er det imidlertid det reelle cepstrumet som er mest brukt. Ofte brukes fourierertransform i stedet for invers fouriertransform ved beregning av cepstrumet.

Utgangspunktet for cepstralanalyse er at et talesignal fremkommer som en foldning mellom et forholdsvis høyfrekvent kildesignal og et mer lavfrekvent filter. I frekvensdomenet blir foldningen til multiplikasjon, og etter logaritmen blir den til addisjon. Vi kan altså se på et logaritmisk talespekter som spekteret til filteret beheftet med høyfrekvent additiv støy fra



Figur 6.1 Cepstralprosessering av et talesignal. Fra (7).

kildesignalet. Ved å ta (invers) fouriertransform av log-spekteret kommer vi til cepstraldomenet der vi kan separere de to bidragene. Kildesignalet ligger i den øvre delen av cepstrumet og filteret i den nedre. Prosessen er vist i figur 6.1.

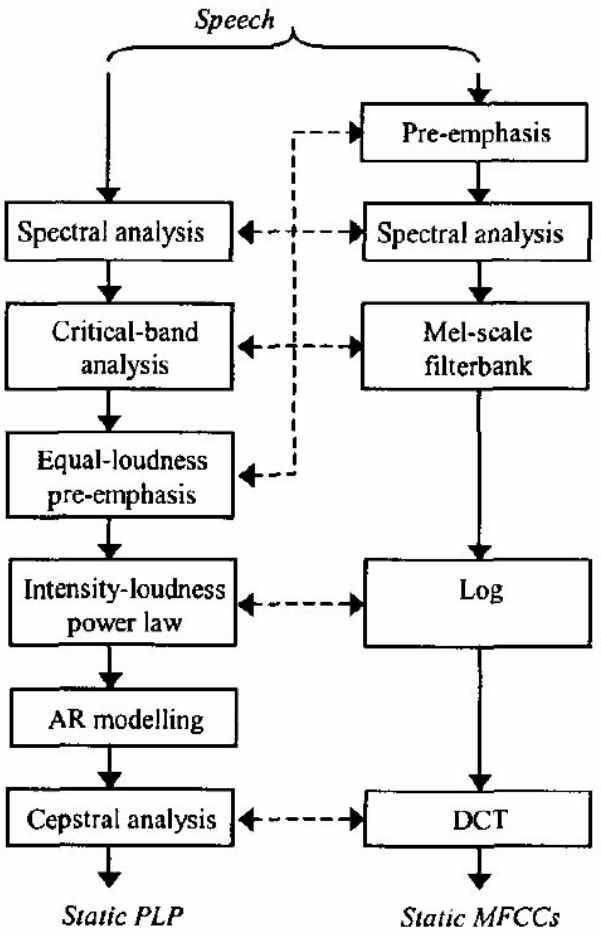
Cepstralanalyse kan altså brukes til å glatte bort den delen av talespekteret som stammer fra stemmebåndene. En annen bruk av cepstrumet kan være å estimere grunnfrekvensen til stemt tale. Denne frekvensen vil vanligvis sees som en enkelttone i cepstrumet.

6.2 Mel-frekvens cepstralkoeffisienter (MFCC)

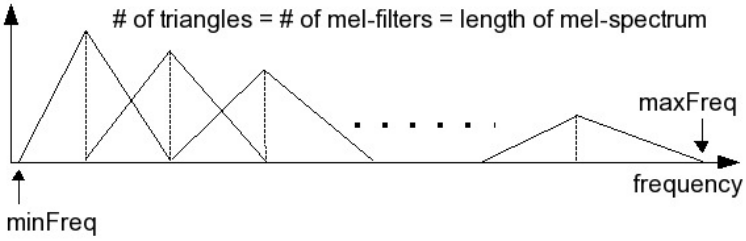
Mel-frekvens cepstralkoeffisienter (MFCC) er muligens den mest brukte metoden for å representere en ramme med taledata for talegjenkjenning. Det som skiller MFCC fra normal cepstralprosessering er at en ikkelineær filterbank som er tilpasset egenskapene til hørselsorganene blir brukt i frekvensplanet. Et flytskjema for beregning av MFCC fins i høyre halvdel av figur 6.2. Første del av beregningene kalles pre-emphasis (15). Særlig for stemt tale er det slik at energien i signalet avtar mot høyere frekvenser. For å motvirke dette brukes ofte et pre-emphasisfilter som i følgende likning

$$H_{pre} = 1 - \alpha z^{-1} \tag{6.2}$$

I likningen er α en konstant som kan varieres mellom 0 og 1, men som typisk settes til 0,95. Dette gir et høypassfilter, som motvirker den spektrale helningen nevnt over. Filteret kan også sees på



Figur 6.2 Flytdiagram for beregning av MFCC og PLP fra (14). Stiplede piler indikerer operasjoner som har omtrent samme formål.



Figur 6.3 En filterbank av mel-filtre. Fra (18).

som en grov etterlikning av hørselens følsomhet, som øker med frekvens i området 1 til 5 kHz (se figur 4.2).

Ved beregning av MFCC brukes en filterbank med M triangulære filter som overlapper som vist i figur 6.3. Toppunktet i hvert filter er uniformt fordelt over mel-frekvensskalaen (se avsnitt 4.3), og hvert filter har samme energi. Antall filter og deres båndbredder varierer. For 16 kHz tale bruker talegjenkjenningsprogrammet Sphinx 40 filter i frekvensområdet 130-6800 Hz (18).

Hvis $X[k]$ er det diskrete korttidsspekteret til en ramme med taledata, og $H_m[k]$ er den diskrete fouriertransformen til filter nummer m i filterbanken, så kan vi beregne log-energien ut av hvert filter som

$$S[m] = \log(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]) \tag{6.3}$$

Man får MFCC-representasjonen av talen ved å ta den diskrete cosinustransformen til $S[m]$.

$$c[n] = \sum_{m=0}^{M-1} S[m] \cos(\frac{\pi n(m - \frac{1}{2})}{M}) \tag{6.4}$$

Cosinustransformen har mange fellestrekk med foruriertransformen. I praksis blir den ofte foretrukket siden den har noe bedre komprimeringsegenskaper enn fouriertransformen. Det vil si at en større del av signalet kan beskrives med færre koeffisienter. I tillegg transformerer den reelle signaler til reelle signaler.

Et typisk talegjenkjenningsprogram vil bruke de 13 første cepstralkoeffisientene som input (10), gjerne sammen med de tilhørende delta- og dobbeltdelta-koeffisientene (se 7). Selv om blant annet antall filter i filterbanken varierer, er dette noe av det nærmeste man kommer en “standard” på området.

6.3 Perseptuell lineærprediksjon (PLP)

Perseptuell lineærprediksjon (PLP) baserer seg på lineærprediksjon og forsøker å modellere en del psykoakustiske konsepter bedre enn MFCC. Et flytskjema for beregning av PLP fins i venstre halvdel av figur 6.2.

Først beregnes FFT for et vindu av talesignalet, vanligvis et Hammingvindu med lengde 20-30 ms. Deretter benyttes en filterbank med filtre jevnt fordelt på bark-skalaen. Filtrene har lik

båndbredde på bark-skalaen og en form som etterlikner de kritiske båndene til basilærmembranen. Videre beregnes energien i signalet fra hvert av filtrene, slik at man får en energikoeffisient fra hvert filter.

For å ta ytterligere hensyn til psykoakustikk justeres energikoeffisientene i henhold til “equal loudness” kurver som beskrevet i avsnitt 4.4. Det vil si at man multipliserer koeffisientene med kurveverdien ved senterfrekvensen til filteret. Etter dette tar man tredjeroten av energikoeffisientene for å etterlikne sammenhengen mellom oppfattet lydstyrke og fysisk effekt.

De resulterende koeffisientene brukes som basis for lineærprediksjon av signalet. Dette gjøres ved hjelp av autokorrelasjonsmetoden, som beskrevet i (8). Lineærprediksjonskoeffisientene brukes til å beregne cepstralkoeffisienter, som utgjør den endelige egenskapsvektoren.

Forskjellen mellom LPC og PLP ligger i forvrengning av frekvens og bruk av filtre som etterlikner de kritiske båndene. Dette fører til en viss spektral glatting som fjerner en del av finstrukturen i signalets spektralinnhold. I tillegg gir det høyere vekting av lavfrekvent innhold i talesignalet, noe som stemmer godt overens med frekvensfølsomheten til den menneskelige hørsel. I forhold til LPC er PLP vesentlig mer støyrobust (11). Dette gjelder både for additiv- og konvolusjonsstøy. Likevel er metodene omtrent like beregningskrevende.

PLP er ikke noe bedre i forhold til MFCC (14). I PLP brukes equal loudness kurver isteden for pre-emphasis, og det brukes en filterbank som er mer tilpasset kritiske bånd enn filterbanken som brukes for MFCC. Likevel ser det ikke ut til at dette gir enn bedre representasjon av talen. Bruk av lineærprediksjon som utgangspunkt for omregning til cepstralkoeffisienter ser heller ikke ut til å ha noen fordel fremfor bruk av diskret cosinustransform.

Det finnes en del egenskapsvektorer som i større grad enn MFCC og PLP prøver å etterligne den menneskelige hørsel. Felles for disse er at de forsøker å modellere overgangen fra mekaniske vibrasjoner til nevral aktivitet i det indre øret. Metodene er generelt dårligere enn MFCC i et støysvakt miljø, men de er mer robuste mot støy. Årsaken er antakelig at de tar hensyn til posisjonen til de dominerende frekvenser i talesignalet. Disse modellene er langt mer beregningskrevende enn MFCC og PLP. Mer informasjon om hørselsbaserte egenskaper kan finnes i (6).

7 TIDSFILTRERING AV EGENSKAPER

De vanligste egenskapsvektorer er basert på korttidsspekter av talesignalet. Det vil si at hver egenskapsvektor beregnes fra rammer av cirka 25 ms med tale. Talesignalet antas å være stasjonært for denne perioden.

I de fleste tilfeller vil et automatisk talegjenkjenningssystem behandle hver egenskapsvektor for seg. Dermed tas det ikke hensyn til effekter som har lengre varighet enn en ramme. Det er mye som tyder på at noe språklig informasjon ligger kodet i hvordan talesignalets parametre endrer seg, ikke bare hvordan de er ved et gitt tidspunkt.

7.1 Tidsderivasjon av egenskaper

En metode for å ta hensyn til hvordan parametre endres er å inkludere den deriverte og eventuelt høyere ordens deriverte av egenskapsvektoren (4, 5). Disse egenskapene kalles ofte med en felles betegnelse for dynamiske egenskaper, og de opprinnelige egenskapene kalles ofte statiske egenskaper. Det er vanlig at førsteordens dynamiske egenskaper, altså estimatet av den tidsderiverte av de statiske egenskapene, kalles for deltaegenskaper. Andreordens dynamiske egenskaper kalles ofte dobbeldelta- eller deltadeltaegenskaper.

I teorien kan man estimere deltaegenskapene som differansen mellom de statiske egenskapene fra to naborammer. Dette viser seg å gi deltaegenskaper med mye støy. Ofte er det mer hensiktsmessig å bruke et glattet estimat som beregnes ved hjelp av flere etterfølgende rammer, for eksempel ved lineærregresjon (15).

En annen metode er å bruke kun to rammer som ikke er naborammer. For eksempel kan man bruke rammene som kommer tre rammer før og tre rammer etter den statiske egenskapen.

Ved bruk av høyereordens dynamiske egenskaper estimeres disse fra de lavere ordens egenskapene. Det er vanlig å bruke samme metode som for beregning av førsteordens dynamiske egenskaper.

I mange tilfeller har det vist seg at resultatene av talegjenkjenningen blir dårligere ved bruk av bare dynamiske egenskaper i stedet for bruk av bare statiske (19). Derfor er det vanlig å bruke statiske egenskaper i kombinasjon med dynamiske, som regel ved å tilføye de dynamiske egenskapene til den statiske egenskapsvektoren. Det er vanlig å bruke like mange delta- og dobbeldeltaegenskaper som statiske egenskaper. Hvis man i utgangspunktet hadde en statisk egenskapsvektor på 13 dimensjoner, som er typisk for MFCC, så vil man få en egenskapsvektor på i alt 39 dimensjoner: 13 statiske-, 13 delta- og 13 dobbeldeltaegenskaper.

Man kan se på utregningen av delta- og dobbeldeltaegenskaper som en filtrering av de statiske egenskapene. Hver dimensjon av den statiske egenskapsvektoren behandles for seg som et tidsvarierende signal. Estimeringen av den deriverte og dobbelderiverte kan implementeres som et lineært tidsinvariant filter. Det kan også være andre filtre som kan være gunstige å bruke på de statiske egenskapene.

7.2 Cepstralmidling (CMN)

Cepstralmidling, eller Cepstral Mean Normalization (CMN), er en teknikk som brukes på statiske egenskaper i cepstraldomenet (13). Teknikken kalles også for Cepstral Mean Subtraction (CMS) eller Cepstral Normalization (CN). Målet er å fjerne konstant eller saktevarierende konvolvert støy. Det vil si at man ønsker å fjerne effekten av for eksempel mikrofon, romakustikk eller liknende.

CMN benytter seg av at konvolusjon er additivt i cepstraldomenet. Det betyr at en kanal med konstant karakteristikk vil opptre som en additiv konstant i cepstraldomenet. Tanken er å midle egenskapene over lang tid og trekke gjennomsnittet fra hver egenskap. Dermed blir man kvitt effekten av den konstante kanalen.

Teknikken fungerer best dersom man kan utføre midlingen over en hel ytring eller mer, altså for

applikasjoner som ikke krever sanntidsytelse. Likevel er det mulig å midle over kortere perioder, slik at teknikken også kan brukes i nær-sanntid.

CMN kan på samme måte som dynamiske egenskaper sees på som en filtrering av de statiske egenskapene.

7.3 Relative Spectral metoden (RASTA)

Relative Spectral metoden (RASTA) (9) er en mer generell filtreringsmetode for statiske egenskaper. Metoden benytter seg av kunnskap om endringshastigheten til talesignalets spektrale komponenter. Det vil si endringen over tid av spektral energi ved forskjellige frekvenser. Denne endringshastigheten kalles ofte for modulasjonsfrekvensen, og det er gjort flere forsøk som viser at mennesket er mest følsom for modulasjonsfrekvenser rundt 4 Hz. De samme undersøkelsene antyder at vanlig tale er modulert med frekvenser i området $1 - 16$ Hz.

En av forklaringene på dette ligger i måten talen produseres på. For å endre uttale må vokaltraktens form endres, og dette krever forflytning av artikulatorene. Det er klart at det er begrenset hvor fort artikulatorene kan bevege seg. Taleproduksjonsapparatet har dermed ikke mulighet til å skape et signal med modulasjonsfrekvens høyere enn cirka 16 Hz.

På bakgrunn av dette kan man båndpassfiltrere de spektrale komponentene av talesignalet. Tanken er at man da får fjernet deler av signalet som i hovedsak inneholder støy og annen irrelevant informasjon.

RASTA metoden brukes i dag hovedsakelig sammen med PLP, og kalles da for RASTA-PLP. Fremgangsmåten for beregning av egenskapsvektorer er den samme som for PLP, bortsett fra at man gjør en filtrering av utgangene fra filterbanken. Utgangssignalet fra hvert av filtrene i filterbanken behandles som et eget tidsvarierende signal. Signalene blir omregnet ved hjelp av en ikkelineær spektral transformasjon, for eksempel logaritme. Deretter filtreres de med et båndpassfilter, før de til slutt transformeres tilbake ved hjelp av den inverse transformasjonen. Ulike transformasjoner og ulike filtre vil ha stor betydning for hvordan RASTA-prosesseringen påvirker egenskapsvektorene.

I prinsippet kan en hvilken som helst ikkelineær transformasjon brukes. Det vanligste er likevel logaritme, fordi man da oppnår en liknende kanalnormaliseringseffekt som ved CMN. Dersom det er stor grad av additiv støy i tillegg til den konvolverte er ikke nødvendigvis logaritmen en god transformasjon å bruke, siden den additive støyen blir signalavhengig etter den logaritmiske transformasjonen. Lin-log RASTA er en metode som forsøker å omgå dette problemet. Der brukes en transformasjon som er lineær for spektrale komponenter med lav energi og logaritmisk for komponenter med høy energi.

På grunn av filtreringen skaper RASTA-metoden avhengighet mellom egenskapsverdier. Dette gjør at metoden ikke er like godt egnet for alle akustiske modeller.

8 OPPSUMMERING

Denne rapporten har tatt for seg metoder for front-end prosessering av talesignaler for automatisk talegjenkjenning. Vi har forsøkt å fokusere på de metoder som er vanligst i dag.

En typisk front-end beregner en representasjon av talesignalet hvert tiende milisekund. Den typiske representasjonen er basert på talesignalets korttidsspekter for en periode på 20-30 ms. Som regel beregnes en egenskapsvektor bestående av 10-15 egenskaper, hvor hver dimensjon i egenskapsvektoren representerer energien i et spektralbånd av talesignalet. Videre er det slik at de fleste egenskaper er beregnet på en slik måte at de etterlikner måten det menneskelige hørselssystemet behandler lydsignaler. Vanligvis inkluderes også et estimat av den deriverte og dobbeltderiverte av egenskapene med hensyn på tid, slik at den endelige egenskapsvektoren har 30-45 dimensjoner.

Den klart mest brukte representasjonen av talesignalet er 13 mel-frekvens cepstralkoeffisienter (MFCC) med første og andre ordens deriverte, gjerne i kombinasjon med cepstralmidling (CMN). Andre mye brukte representasjoner inkluderer perseptuell lineærprediksjon (PLP) og lineærpredikterte cepstralkoeffisienter (LPCC).

Litteratur

(1) Bogart B, Healy M og Tukey J (1963): The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. I Rosenblat M, redaktør, *Proceedings of the Symposium on Time Series Analysis*, side 209–243. Wiley, N.Y.

(2) Fant G (1970): *Acoustic theory of speech production*. Walter de Gruyter, inc.

(3) Fletcher H (1938): Loudness, masking and their relationship to the hearing process and the problem of noise measurement. *Acoustical Society of America*, 9:275–293.

(4) Furui S (1981): Comparisoon of Speaker Recognition Methods Using Statistical Features and Dynamic Features. *IEEE Trans. Acoustics, Speech and Signal Processing*, 29(3).

(5) Furui S (1986): Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. Acoustics, Speech and Signal Processing*, 34(1).

(6) Gajic B (2003): Auditory Based Methods for Robust Feature Extraction. I Espvik O, redaktør, *Telektronikk: Spoken Language Technology in Telecommunications*, bind 99, side 45–58. Telenor Communications AS.

(7) Harrington J og Cassidy S (1999): *Techniques in Speech Acoustics*. Kluwer Academic Press.

(8) Hermansky H (1990): Perceptual Linear Predictive (PLP) analysis of speech. *Acoustical Society of America*, 87:1738–1752.

(9) Hermansky H (1994): RASTA processing of speech. *IEEE Trans. Speech and Audio Processing*, 2(4).

(10) Huang X, Acero A og Hon H (2001): *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice-Hall PTR.

(11) Jankowski C R, Vo H D H og Lippmann R P (1995): A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. Speech Audio Processing*, 3(4):286–293.

(12) Lipshitz S P, Pocock M og Vanderkooy J (1982): On the Audibility of Midrange Phase Distortion in Audio Systems. *J. Audio Eng. Soc.*, 30(9):580–595.

(13) Liu F H, Stern R M, Huang X og Acero A (1993): Efficient Cepstral Normalization for Robust Speech Recognition. *Proc. of the sixth ARPA Workshop on Human Language Technology*.

(14) Milner B (2002): A Comparison of Front-end Configurations for Robust Speech Recognition. I *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, bind 1, side 797–800.

(15) Picone J W (1993): Signal Modeling Techniques in Speech Recognition. *Proc. IEEE*, 81(9).

(16) Proakis J G og Manolakis D G (1996): *Digital Signal Processing: Principles, Algorithms and Applications*. Prentice-Hall International, inc.

(17) Rabiner L R og Schafer R (1978): *Digital Processing of Speech Signals*. Signal Processing. Prentice Hall, Englewood Cliffs, NJ.

(18) The CMU Sphinx Group Open Source Speech Recognition Engines. http://cmusphinx.sourceforge.net/sphinx4/.

(19) Wrede B og Fink G A (2003): What is in the Dynamic Features: Analysis of the Derivative of log-mel-spectra. *Proc. Int. Congress of Phonetic Science*.