Institutt for elektronikk og telekommunikasjon

Eksamensoppgave i TTT4185 Taleteknologi

Faglig kontakt under eksamen: Torbjørn Svendsen
Tlf.: +47 930 80 477

Eksamensdato: Torsdag 4. desember 2014

Eksamenstid (fra - til): 09.00 - 13.00

Hjelpemiddelkode/tillatte hjelpemidler: **C** – Spesifiserte trykte og håndskrevne hjelpemidler tillatt. Bestemt, enkel kalkulator tillatt.

Annen informasjon:

- Eksamen består av 3 oppgaver der
  - oppgave 1 omhandler taleanalyse
  - oppgave 2 omhandler talegjenkjenning
  - oppgave 3 omhandler talesyntese
- Alle deloppgaver teller likt
- Alle oppgavene skal besvares
- Sensurfrist er 3 uker etter eksamensdato.

Målform/språk: Norsk - bokmål

Totalt antall sider: 8

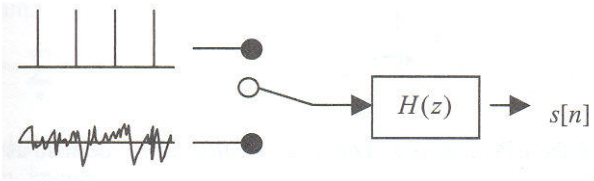
Herav, antall vedleggsider:

Kontrollert av:

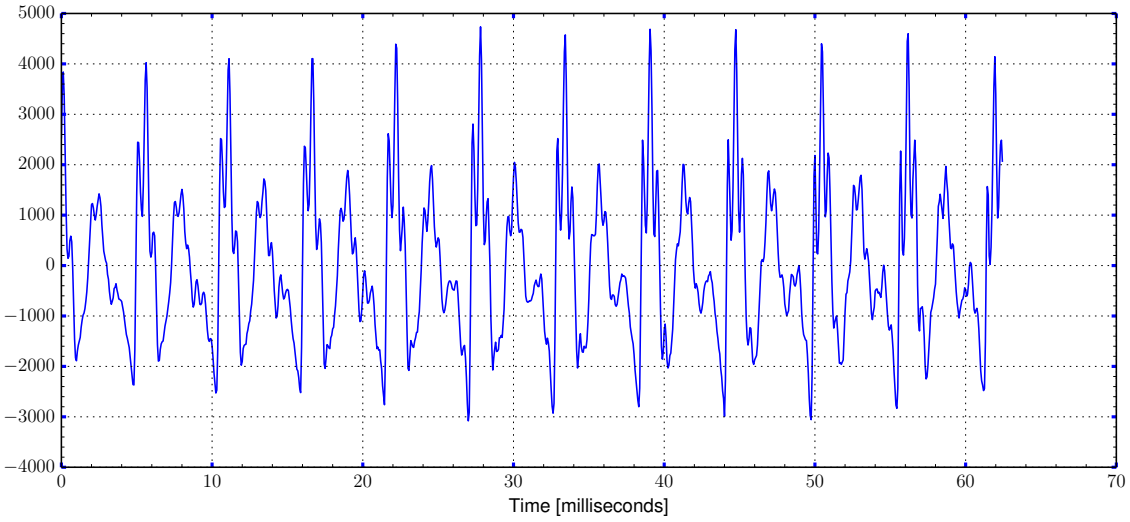
Dato Signatur

Oppgave 1

1a) Skisser kilde-filter modellen for taleproduksjon, og grei ut om de forskjellige delene av modellen og hva de representerer fysisk.



Skissen over er en kilde-filter modell. Til venstre finner vi eksitasjonen, som enten er i form av et pulstog fra stemmebåndene eller turbulent luft. Til høyre finner vi filteret, som er en representasjon av taleorganet.



Figur 1: Utsnitt av bølgeformen til en vokal

1b) I figurene 1 og 2 er hhv. bølgeformen og spektrumet til samme vokal representert grafisk. Finn pitsjen ved hjelp av de to figurene og beskriv framgangsmåten. (De to estimatene vil ikke nødvendigvis ble eksakt like).

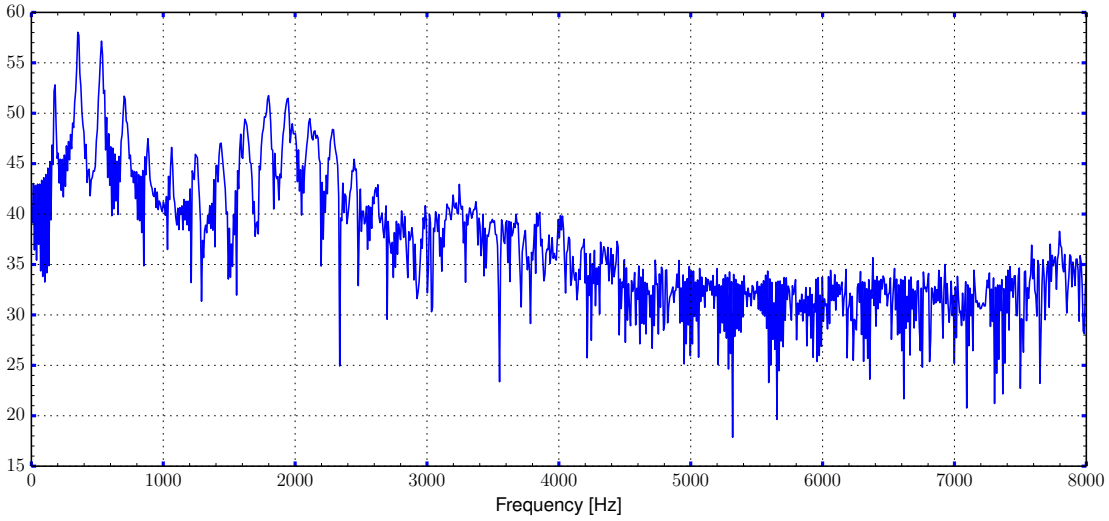
Fra bølgeformen i figur 1 ser vi at vi har 11 pulser i løpet av ca 62 millisekunder. Dette gir en pitsj på $\frac{11}{0.062s} \approx 177Hz$.

Fra spektrogrammet ser vi at det typiske pitsjmønsteret har 8 perioder i området 0-1400 Hz. Dette gir en pitsj på $\frac{1400Hz}{8} \approx 175Hz$

1c) Estimer de tre første formantene ved hjelp av figur 2. Hvorfor er formantenes posisjon og form viktig for klassifisering?

Formantene finner vi på ca 400, 1800 og 3300 Hz.

Posisjonen og formen til formantene reflekterer konfigurasjonen på taleorganet og dermed hvilket fonem som uttales.



Figur 2: Spektrumet til et utsnitt av en vokal

1d) Beskriv hvordan omhyllingskurven til spektrumet kan estimeres ved hjelp av lineær prediksjon. Hvilken orden bør lineær prediksjonsfilteret minst ha for å beskrive posisjonen til tre formanter?

Basert på kilde-filter modellen antar vi at signalet $x(n)$ er av formen

$$x(n) = \sum_{i=1}^{p} h_i x(n-i) + e(n) \tag{1}$$

der h_i er filteret og $e(n)$ er eksitasjonen. Et lineært predisjonsfilter estimerer h_i ved å minimaliserer predisjonsfeilen

$$\hat{h} = \underset{h}{\operatorname{argmax}}\, E = \underset{h}{\operatorname{argmax}} \sum_{n} \hat{e}^2(n) \tag{2}$$

der

$$\begin{aligned} \hat{e}(n) &= x(n) - \hat{x}(n) \\ &= x(n) - \sum_{i=1}^{p} h_i x(n-i) \end{aligned} \tag{3}$$

Man trenger to koeffisienter for å beskrive en harmonisk resonans, dvs. en topp"i spekteret. Her trenger man mao. seks koeffisienter for å beskrive tre topper.

1e) Det reelle cepstrumet til et signal $x(n)$ er gitt ved

$$c(k) = \text{IDFT}\left\{\log|X(\omega)|\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)| e^{i\omega k} d\omega \tag{4}$$

der $X(\omega)$ er den Fouriertransformerte til $x(n)$. Vis at cepstrumet til et signal $y(n) = h(n) \star x(n)$, der $\star$ indikerer foldning, er summen av cepstrumene til hhv. $h(n)$ og $x(n)$.

$$
\begin{aligned}
c_y(k) &= \text{IDFT}\left\{\log|Y(\omega)|\right\} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|Y(\omega)| e^{i\omega k} d\omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)H(\omega)| e^{i\omega k} d\omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\log|X(\omega)| + \log|H(\omega)|\right) e^{i\omega k} d\omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(\omega)| e^{i\omega k} d\omega + \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|H(\omega)| e^{i\omega k} d\omega \\
&= c_x(k) + c_h(k)
\end{aligned}
\tag{5}
$$

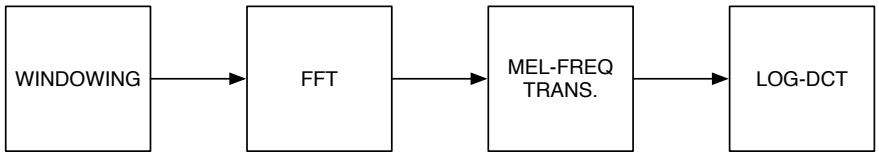
1f) Forklar hvordan man kan bruke cepstrumet til å estimere omhyllingskurven til signalet i figur 2.

I cepstral-domenet er kanalen additiv til signalet. Signalene er separable siden kanalen stot sett er glatt og saktevarierende og derfor representert av $c(k)$ for lave k, mens eksitasjonen x er hurtigvarierende og representeres av høye k. Ved å sette

$$
\hat{c}(k) = \begin{cases} c(k), & k < K \\ 0, & K \leq k \end{cases}
\tag{6}
$$

og deretter gjenvinne log-spekteret fra $\hat{c}(k)$ kan vi finne omhyllingskurven.

1g) Den mest brukte egenskapsuttrekningen for bruk i talegjenkjenning er *Mel Frequency Cepstral Coefficients* (MFCC). Tegn blokkskjema som viser hvordan man går fra et punktprøvet talesignal til en sekvens av MFCC-vektorer.

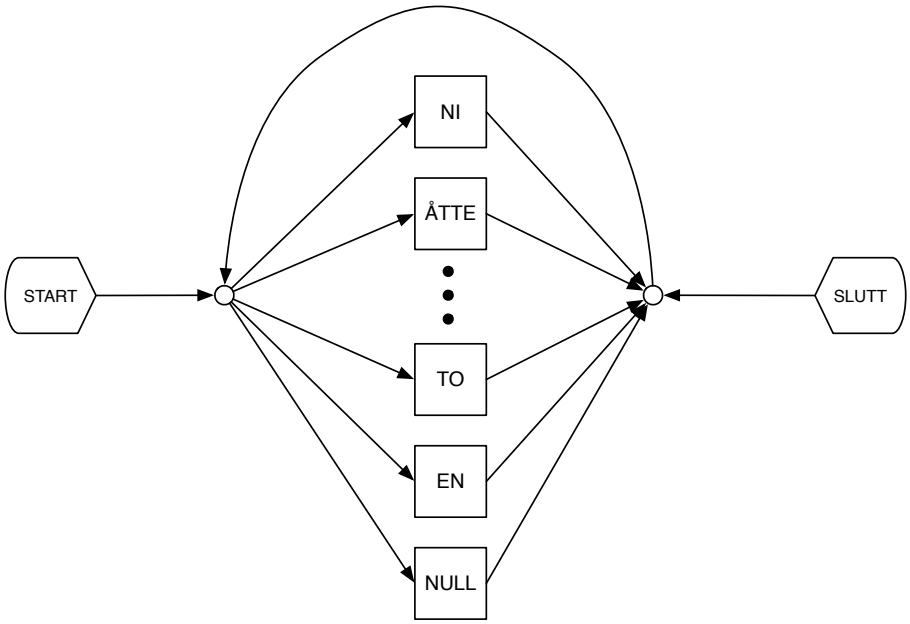


Prosesseringen skjer som følger:

- WINDOWING: Talesignalet deles opp i små, overlappende blokker som pålegges et vindu (typisk Hamming).
- FFT: Blokken transformeres til frekvensdomenet
- MEL-FREQ: Energien i overlappende delbånd av økende bredde beregnes. Delbåndenes posisjon og bredde er gitt av Mel-skalaen, en perseptuell frekvensskala.

LOG-DCT: Energien i delbåndene representeres ved sin log-verdi og transformeres vha. en diskret cosinus transform. Dette er ment å dekorrelere egenskapsvektoren.

Oppgave 2

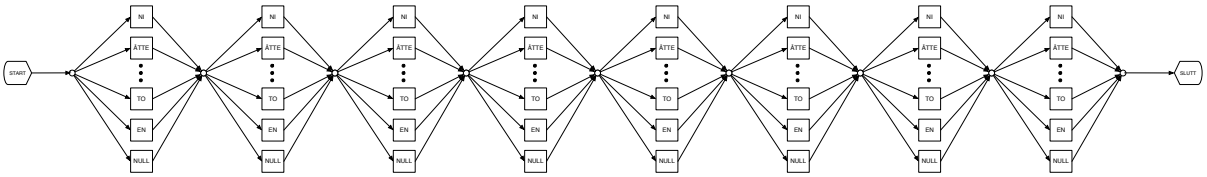


Figur 3: Enkel grammatikk for gjenkjenning av telefonnummer

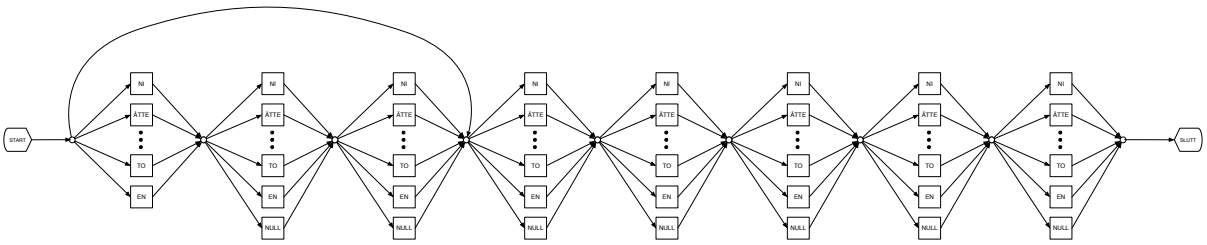
En ingeniør har fått som oppgave å designe et applikasjon som skal lese inn telefonnummer vha. talegjenkjenning. Et første utkast til grammatikk er gitt i figur 3. Merkene `START` og `SLUTT` er ikke ord som kan gjenkjennes, men tilstander som indikerer starten og slutten på ytringen.

2a) Vil denne grammatikken alltid gi resultater som er meningsfulle telefonnnumer? Utdyp svaret. Kom med forslag til endringer av grammatikken som alltid gir ut et åttesifret nummer.

Grammatikken vil kunne gi alle mulige sekvenser for alle mulige sekvenslengder N, og vil derfor også tillate tallstrenger som ikke er gyldige telefonnummer.



2b) Ingeniøren blir bedt om å forbedre grammatikken ytterligere slik at den nå kan gi ut enten åttesifrede tall som *ikke* kan starte med null, eller femsifrede tall som kan starte med null. Skisser forslag til en slik grammatikk.



2c) Grei ut om generelle fordeler og ulemper ved helord, stavelser, fonemer og trifoner som basisenheter for talegjengkjenning. Talegjenkjenneren skal virke i hele landet og gjenkjenne alle uttaler av tallene. Det blir bestemt at talegjenkjenneren skal basere seg på helordsmodeller. Diskuter dette valget. Hva er en alternativ løsning?

There are mainly three characteristics we want a speech unit to have: accuracy, trainability and generalizability.

- Words
  - High accuracy, since the all coarticulation effects are internal to the word.
  - Low trainability since the number of words are is very high.
  - Low generalizability as a new data needs to be collected and models trained if new words are to added
  - Suited for tasks with small vocabularies
- Syllables
  - High accuracy, since the strongest word-internal coarticulation effects are internal to the syllables.
  - Low trainability since the number of syllables that can occur (especially if proper names are included) is very high.
  - Good generalizability.
  - Suited for tasks with abundant training data.
- Phones are the most common speech unit in current systems, and can be further divided into two types:
  - Context independent phones
    * Low accuracy since coarticulation effects are not modeled.
    * Very high trainability since the number of phonemes is low (40-50).
    * Very high generalizability.
    * Suited for tasks with little available training data and little information about the kinds of input the system will receive.
  - Context dependent phones
    * High accuracy since coarticulation effects over triphones are modelled.
    * Low trainability since the number is very high. (However, clustering states where different contexts give rise to similar coarticulation effects can reduce this number considerably.)
    * Good generalizability.
    * Well suited for all tasks where there is sufficient data available to train them.

Vokabularet er lite, så i utgangspunktet virker helordsmodeller som et godt valg. Siden systemet skal virke over hele landet vil man imidlertid måtte hente inn og trene svært mange unntak. Det kan derfor være bedre å bruke en delordsmodell, og heller legge til alternative talemåter i uttaleleksikonet.

2d) Et problem som oppstår under bruk er at systemet gir ut et telefonnnumer uansett hva som brukeren sier. Ingeniøren lager seg to statistiske modeller, $P(X|\text{nummer})$ og $P(X|\text{annet})$, der X er ytringen. Han vet dessuten etter å ha sett på systemloggene at 95% av ytringene er nummer, mens de resterende 5% er annet. Bruk disse opplysningene til å sette opp et

uttrykk for $P(\text{nummer}|X)$. Hvis $P(X|\text{nummer}) = 0.01$, og $P(X|\text{annet}) = 0.05$ – skal da ytringen forkastes eller ikke?

$$P(\text{nummer}|X) = \frac{P(X|\text{nummer})P(\text{nummer})}{P(X)} \tag{7}$$

Setter inn tallverdiene:

$$\begin{aligned} P(\text{nummer}|X) &= \frac{0.01 \times 0.95}{P(X)} = \frac{0.0095}{P(X)} \\ P(\text{annet}|X) &= \frac{0.05 \times 0.05}{P(X)} = \frac{0.0025}{P(X)} \end{aligned} \tag{8}$$

Siden $P(X)$ er felles for begge sannsynlighetene slipper vi å beregne denne, og vi antar at det var et nummer som ble mottatt.

2e) Modellen for $P(X|\text{annet})$ er en Gaussisk blandingsfordeling (GMM). Hvordan vil du gå fram for å trene en slik modell, gitt et sett med N datavektorer som representerer annet enn telefonnnummer? Gitt at systemet er basert på egenskapsvektorer med 39 komponenter og at ingeniøren valgte å bruke 64 blandingskomponenter og diagonale kovariansmatriser – hvor mange frie parametre vil måtte estimeres i denne GMMen?

Gaussisk blandingsfordelinger (GMM) er et eksempel på modeller som har "missing data", i dette tilfellet identiteten til blandingskomponenten som hver enkelt vektor ble trukket fra. Vi kan bruke EM-algoritmen til å estimere parametrene.

Antall frie parametre blir:

- Komponentvekter: 63 (64 godtaes også, men 63 er riktig svar da alle 64 må summere til én)
- Middelverdivektorer: $64 \times 39 = 2496$
- Kovariansmatriser, diagonale: $64 \times 39 = 2496$

Totalt: $2 \times 2496 + 63 = 5055$ (5056 godtaes også)

Oppgave 3

3a) Hvilke fire hovedblokker utgjør et talesyntesesystem? Beskriv kort funksjonaliteten til hver av blokkene.

- Text analysis: text normalization; analysis of document structure, linguistic analysis
  Output: tagged text
- Phonemic analysis : homograph disambiguation, morphological analysis, letter-to-sound mapping
  Output: tagged phone sequence
- Prosodic analysis: intonation; duration; volume
  Output: control sequence, tagged phones
- Speech synthesis: voice rendering
  Output: synthetic speech

3b) Forklar fenomenet *Large Number of Rare Events* og hva det har å si for datadrevet skjøtesyntese.

- Large number of units with small probability of occurrence
- If database units are selected randomly, the probability of encountering a unit not in the database approaches certainty for a small sequence of randomly selected sentences.
- Unit inventory must be chosen with care
- Fall-back solutions must exist for non-covered units

3c) Beskriv PSOLA-algoritmen og forklar hvordan den brukes i et system basert på difonsyntese. Hvorfor kan PSOLA-algoritmen også være nyttig i et system basert på datadrevet skjøtesyntese?

PSOLA can modify pitch and duration. The starting point for PSOLA is having accurate estimates of the fundamental frequency in the voiced parts of the speech signal. For every vocal pulse you can then construct a waveform segment centered around the pulse. The segments are attenuated towards the end points and typically extend over two fundamental periods (2T0). Through the additive combination of partially overlapping segments, where the degree of overlap is such that the distance between successive vocal pulses is equivalent to the desired fundamental frequency, a voice signal with the same spectral envelope but a new fundamental frequency is constructed. By repeating or skipping segments from the original signal, the manipulation can be done without changing the duration of the signal.

I difonsyntese brukes PSOLA alltid til å glatte overgangene mellom to difoner.

I et system basert på datadrevet skjøtesyntese kan PSOLA brukes til å kompensere for at man ikke finner helt riktig enhet.

3d) Grei ut om fordeler og ulemper med å benytte korte taleenheter som fonemer eller difoner, kontra lengre enheter som ord og fraser i konkatenativ talesyntese.

Lengre enheter fører til bedre kvalitet, men krever mer lagringsplass og er mer kontekstavhengige. Mest brukt for begrensede domenespesifikke systemer.

Små enheter er mer trenbare, krever mindre lagringsplass og er svært fleksible.