THE CIRCUIT BREAKERS SIN INNLEVERING AV

ICA04

Modell for digitalisering av analoge lydsignaler samt naturlig språk gjenkjenning

Modell for digitalisering av analoge lydsignaler

Det kan virke forhåndsvis enkelt finne veien fra analogt stemmesignal til digitalt forstått materiale. Det viser seg ved nærmere inspeksjon at det er utrolig komplisert.

Når vi mennesker snakker, bruker vi et sett med modulatorer for å manipulere strømmen av luft vi presser ut av lungene våre. Disse modulatorene er stemmebånd, tunge, leppene og kjeven. Når vi påfører disse "modifikasjonene" til luftstrømmen vil den resonnere på forskjellig måte, og på den måten skape en kontinuerlig (ikke diskret) analog lydkrekvens. Lyd som stammer fra "stemmen" vår kan også beskrives som hvor ofte variasjoner i luftrykket forekommer.

For at en PC skal kunne tolke hva vi sier må vi først sørge for å "snakke datamaskinens språk" - nemlig digitalt. Det analoge signalet vi har sendt ut av munnen vår må gjennom en digitaliseringsprosess. Etter at lyden har truffet membranen i en mikrofon, vil en analog-til-digital konverteringsalgoritme kode hver seksjon av lydbølgen til et set med bits. For at dette skal bli gjengitt på best mulig måte bruker man fysikeren Harry Nyquist sin teori om at sampling av lyd best gjennomføres ved dobbling av høyeste input frekvens. Ved en input på 4kHz bruker vi da en samplingfrekvens på 8kHz. Da vil det analoge signalet bli kodet til 8-bit (1byte) hvert 125. microsekund (1/8000 av et sekund).

Ved at et analogt signal er samplet med en hyppighet på 8000 ganger i sekunder bruker vi en del plass. Her finnes det spesielle kompresseringsalgoritmer som er utviklet av både linguister og biologer side om side - disse vil minimere plassen det digitale signalet tar opp under overføring. Når vi snakker, som nevnt over i artikkelen, endrer vi den resonante frekvensen i lydbølgene - ellers kalt de formante frekvensene. Hvis man tenker på halsen som en transportsmedium så blir frekvensen påvirket av hvordan den beveger seg gjennom dette mediumet. Lengden frekvensen må reise, samt rommet den må reise gjennom, som vi begge styrer med modulatorene våre, påvirker resultatet. På denne måten kan man relatere formanter med hva som skjer i tranittmediumet (halsen++). Det

vil si at man vet at enkelte lyder (formanter) oppstår på forskjellige steder. Det er for eksempel enighet om at F1, den første formanten har å gjøre med tungens høyde. Jo høyere tungen er, desto lavere er F1.

Nå signalet er digitalisert og komprimert kan vi begynne å leke oss med informasjonen. Nå forsår PC'en at vi snakker samme språk, og informasjonen tar lite nok plass til at den blir effektiv å sende og motta. I det neste steget har vi valgt å benytte oss av wit.ai, en tjeneste for utviklere som dekoder stemmefiler og hjelper deg med å trekke ut mening fra lyd eller tekst, og gi deg et mye mer oversiktlig JSON objekt. wit.ai benytter seg av flere forskjellige metoder for å "dra" mening ut av filene vi sender. Den viktigste er statistisk analyse, hvor filen vi sendte blir sammenlignet med over 2,5 milliarder andre filer som ligger inne i databasen til wit.ai. Her er det flere algoritmer som "kverner" dataen, bla.

- Mønsteranalyse
    - Består av en liten ordbok, en variabel sjekkes opp mot denne
- Mønster og bemekelsesanalyse
    - Mye mer komplisert en konvensjonell mønsteranalyse.
- Statistisk analyse
    - Desto mer data som er tilgjengelig, desto mer effektivt. Den tar det den tror du sa og matcher det mot en database for å sjekke sannsynligheten for hvilken rekkefølge ordene kommer i
- Kunstige nervenettverk
    - Kan basere det du sier på noe du har sagt før. Kan lære seg dialekten din.