Akustisk modell for automastisk stemmegjenkjenning

Den akustiske modellen representerer forholdet mellom lydsignaler og fonemer og andre språkelementer som utgjør vår talemåte.   
Fonemer er lyder i ord vårt språk og tale består av. Det går ikke på hvordan ord blir stavet, men på utale. Den akustiske modellen lærer ord fra en rekke lydinnspillinger av vanlig tale og skriftlig fremstilling av det som blir sagt. Et program beregner så hvilke ord som blir sagt utifra lydinnspillingene. Det er omkring 40 lyder i det engelske språket. Lyden ”F” kan staves på flere forskjellige måter, f, ph, gh f. eksempel.

Det første bildet viser et godt eksempel på ulike lyder, og at de ikke har noe sammenheng med andre ord. Den akustiske modellen må derfor lære seg forskjellen på lydsegmenter for å unngå å oversette feil. Ved å for eksempel spille inn ordet «enough» gjentatte ganger sammen med en skrevet fremstilling, vil modellen lære å gjenkjenne ordet og stave det riktig.

I moderne talegjenkjenningssystemer brukes den akustiske og en språk-modell for å beregne egenskapene ved tale. Den akustiske modellen beregner forholdet mellom lydsignalet og lydene i språket, mens språk-modellen setter sammen ord til setninger, og skiller mellom ord og utrykk som høres likt ut.

«Wreck a nice beach» og «recognize speech» er to setninger som høres likt, men betydningen er selvsagt veldig forskjellig. Det samme gjelder «Ice cream» og «I scream». Derfor må også modellen lære å gjenkjenne individuelle ord i setninger. Dette gjøres ved å analysere hvor lang tid og hvor vi legger trykk på ordene når vi sier dem.

Det andre bildet viser hvordan et talegjenkjenningssystem tar inn lydfrekvens, og bruker modellene for å kode lyden for å sette sammen en fullstendig setning

Lyden kan bli kodet i forksjellig frekvenser, hvor mange hertz per sekund, og forskjellig bits per sekund. Det beste er å bruke samme akustiske modell som har ”trent” på samme frekvens og bits per samling som talen den skal gjenkjenne.

En god regel er at talegjenkjenningen virker best hvis modellen er ”trent” med lyddata med høy frekvens og bits per sekund, men igjen, for høye verdier vil gjøre at prosessen tar tid. Standard for akustiske modeller er 16 kHz med 16 bits per frekvens, 256 000 bits per sekund.