Användning av läppläsning i framtida hörapparater

Viktor Andersson, Nelly Ostréus

Det vanligaste sättet att kommunicera är via tal. För att förstå tal och för att kunna hänga med i ett samtal används främst hörseln, men även synen spelar roll. Att använda visuella intryck för att förstå tal kallas för läppläsning. Eftersom mer än en miljard människor har en hörselnedsättning, så är läppläsning en stor del i många personers vardag.

I en stökig miljö där flera personer pratar samtidigt, har en person med en hörselnedsättning ofta svårt att fokusera hörseln på en person. Detta sker i vardagen vid till exempel restaurangbesök eller på fester. Ett problem med dagens hörapparater är att de inte hjälper användaren att fokusera på personen som denne vill lyssna på. Det här examensarbetet är gjort i samarbete med hörapparatsföretaget Oticon för att kunna utveckla bättre hörapparater som tillåter användaren att fokusera på en röst i en rörig miljö.

Mer specifikt har det i det här examensarbetet undersökts om man kan förutsäga om en person pratar eller inte pratar i en video när man inte har tillgång till ljudet på videon. Alla videor som har använts har sett ut som i Figur 1, där en skådespelare framför en monolog och de andra två skådespelarna framför en dialog.

Varje skådespelare i videon hade en mikrofon som spelade in ljudet. Skådespelarnas tal har sedan analyserats och klassificerats som tal eller icke-tal. Olika kännetecken användes på varje person i videon för att klassificera om personen pratade eller ej. I vårt arbete använde vi arean av munnen och sträckan mellan över- och underläppen som kännetecken. Vi använde sedan en maskininlärningsalgoritm för att förutsäga om en person på videon pratar eller inte pratar.

Vi hittade ingen korrelation i monologerna mellan kännetecknen från videorna och om skådespelaren pratar eller inte. I dialogerna var det istället väldigt blandade resultat för de olika videorna. För en del videor hittade vi ett tydligt samband och algoritmen kunde i relativt hög grad förutsäga om personen pratade eller inte, medan andra videor inte visade något samband alls. Bäst resultat uppnåddes när vi använde ett mått på hur mycket munarean varierade under 0,25 sekunder för att förutsäga om personen talade eller inte.

Figur 1. En bild från en video med en monolog framförd av skådespelaren till vänster och en dialog framförd av skådespelaren till höger.

Det finns flera förbättringsområden för att uppnå ännu bättre resultat. Det finns flera forskningsgrupper som menar på att man behöver studera hela ansiktet och inte bara munnen för att kunna avgöra om någon pratar eller inte. Det hade därför varit intressant att vidareutveckla projektet genom att använda kännetecken från hela ansiktet och inte bara från munnen. Det finns också flera fungerande algoritmer inom området automatiserad läppläsning, men som är betydligt mer komplexa än de som använts i det här examensarbetet. En naturlig fortsättning på projektet hade också kunnat vara att testa så kallade djupinlärningsmetoder för att bättre förutsäga om en person pratar eller inte.