Att lägga ord i munnen på varandra: Undersökning av den inferentiella modellen av talproduktion

Valery Nkenguruke
va2518nk-s@student.lu.se

I denna artikel rapporteras det om ett pågående forskningsprojekt ämnat att utforska den så kallade ”inferential model of speech processing” med hjälp av det så kallade Real-time Speech Exchange (RSE)-systemet. Projektet syftar till att förstå rollen av auditiv feedback i talmonitorering och uppfattningen av egen handlingskraft i individuella och sociala konversationsmiljöer. Här kommer jag att presentera de preliminära studier (pilotstudier) som jag och en annan forskningsassistent hade som uppgift att utföra. Dessa preliminära studier har haft varierande experimentella förhållanden, allt från strukturerade minnesuppgifter till fria konversationsinteraktioner. Forskningen integrerar innovativa metoder för att undersöka den dynamiska och kontextberoende karaktären av talproduktion och monitorering, och utmanar traditionella psykolingvistiska modeller som anger fördefinierade språkliga avsikter.

1 Introduktion/Bakgrund

Traditionellt är talproduktion och monitorering förstådd genom den så kallade komparatormodellen. Enligt den antas tal börja med en klar definierad preverbal intention, vilket ger ett riktmärke för själv-monitorering och en känsla av agens för ens uttalanden. Denna förhärskande modell antar att talmonitorering sker genom att jämföra det avsedda meddelandet med den faktiska talade utsignalen för att upptäcka och korrigera eventuella avvikelser (Levelt et al., 1999 & Pickering & Garrod, 2013). Den auditiva feedbacken används då enbart till att felkorrigera den talade utsignalen.

I kontrast föreslår den inferentiella modellen att talare ofta börjar prata utan ett medvetet plan för vad de kommer att säga, och dynamiskt konstruerar talets mening baserat på pågående auditiv feedback och sammanhang. Modellen antyder alltså att den auditiva feedbacken, bland annat, utöver dess roll i att upptäcka fel i det uttalade signalen, används för att sluta sig till talares faktiska intention. Denna modell har fått stöd genom experiment som använt tekniken Real-time Speech Exchange (RSE), som visade att talare ofta inte upptäckte manipulationer i sin auditiva feedback, alltså att de inte upptäckte avvikelser mellan vad de sade och vad de hörde sig själv säga (Lind et al., 2014). Detta tyder på en flexibel och adaptiv talmonitoreringsprocess än vad tidigare modeller har föreslagit.

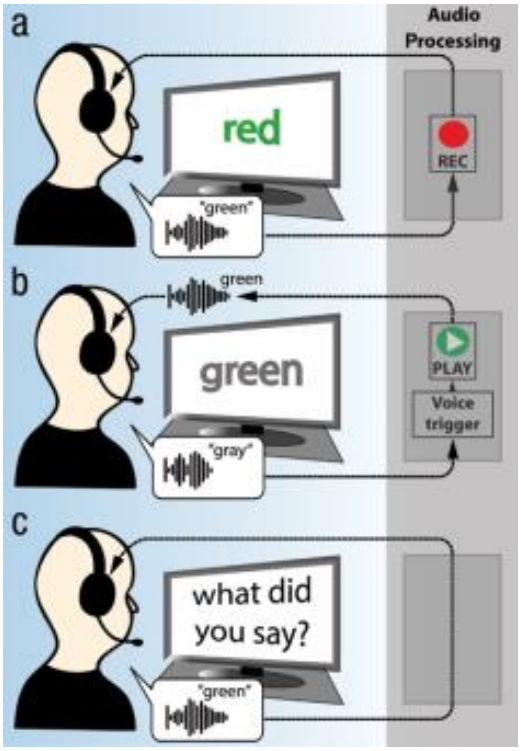

Fig. 1. Illustration av en tidigare experimentell procedur där deltagarna utförde ett Stroop-test, där de ombads att namnge teckensnittsfärgen för varje ord som presenterades på skärmen men kunde i stället höra det färgord som namngavs av bokstäverna.

Den inferentiella modellen betonar alltså att talproduktion och meningsskapande är en dynamisk och kontinuerlig process som starkt påverkas av kontextuella signaler och auditiv feedback, snarare än att vara styrd av förutbestämda intentioner. Detta forskningsprojekt bygger på dessa insikter för att ytterligare utvidga vår förståelse av talproduktionsmekanismer och bidra till teoretiska framsteg inom kognitiv vetenskap när det gäller hur tal övervakas och justeras i realtid samt validera den inferentiella modellen.

2 Metod

Datainsamlingsmetoder

Experimenten använde både kvalitativa och kvantitativa datainsamlingstekniker, inklusive ljudinspelning av deltagarnas svar, och post-test intervjuer för att samla subjektiva intryck av upplevelsen. Den primära metoden för datainsamling involverade användningen av RSE-systemet. Försökspersonen ägnade sig åt uppgifter utformade för att framkalla tal under kontrollerade förhållanden, som att beskriva bilder eller återkalla ordsekvenser.

RSE-Tekniken är en hemmabyggd mjukvara, bestående av ett headset med hög passiv ljuddämpning. Under

manipulerade trials aktiverar försöksledaren en rösttrigger, och när mikrofonsignalen överskrider en förinställd amplitud, ersätts ett tidigare inspelat ord av det uttalade ordet i den auditiva återkopplingen samtidigt som ljudet av deltagarens faktiska uttalande blockerades. Alltså kan försöksledare vid ett önskat tillfälle med en knapp göra så att försökspersonen hör sig själv säga något de sa vid ett tidigare tillfälle. Den som sköter RSE måste sitta fokuserad, spela in ett visst ord genom att trycka på en viss knapp för att sedan byta ut den mot ett annat.

En av de främsta utmaningarna i vårt projekt var att få bra tajming i utbyten med RSE. Framgången för den auditiva manipulationen berodde kritiskt på precisionen av denna tajming; eventuella feltajmingar mellan vad deltagarna sa och vad de hörde kunde lätt upptäckas, vilket undergrävde experimentets giltighet.

Vi använde oss av psychopy mjukvaran som är ett mjukvarupaket känt för sin användning inom neuro- och kognitionsvetenskap och experimentell psykologisk forskning för att presentera visuella stimuli av våra experiment. Dessutom använde vi paper för att presentera bilder på olika saker ämnad att få i gång tankar eller associationer

Analysmetoder

De inspelade data analyserades med hjälp av en programvara för att analysera talinspelningarna som heter Praat, ett verktyg som är allmänt känt för sina robusta funktioner inom fonetisk analys. Denna programvara gjorde det möjligt för oss att exakt mäta variabler som taltiming, tonhöjd och volym, som var avgörande för att bedöma effekterna av auditiv manipulationer.

3 Forskningsprocess

Detta projekt innefattade genomförandet av en serie pilotstudier för att systematiskt undersöka förutsägelserna inom den inferentiella modellen. Varje studie designades för att undersöka olika aspekter av talmonitorering och känslan av agens under förhållanden med manipulerad auditiv feedback.

Projektet inleddes för min del med diskussioner med min handledare för att fastställa en tydlig riktning. Vi utarbetade ett preliminärt schema, och jag tilldelades relevanta artiklar att läsa för att bygga en omfattande förståelse av befintlig forskning. Denna litteraturgenomgång var avgörande för att grunda mitt experimentella tillvägagångssätt i etablerad forskning. Jag lärde mig även hur RSE-systemet fungerar genom övningar med mig själv, min handledare och kurskamrater, för att bli bekväm med systemet som var centralt för experimenten.

De efterföljande pilotstudierna genomfördes på Humanities Lab vid SOL. Jag ansvarade, ibland tillsammans med en annan forskningsassistent, för att rekrytera deltagare, genomföra sessionerna, samla in data och hantera tekniska inställningar, inklusive RSE-systemet, samt analysera data. Varje studie inleddes med teknisk uppställning, som inkluderade bland annat två datorer (en för RSE-mjukvaran och en för Psychopy-programmet där experimentet kördes), en datorskärm för försökspersonen, minst en specialbyggda headset med ljuddämpning, hörlurar (försöksledare hade vanliga hörlurar), samt sändare och mottagare för mikrofonerna.

Nästa steg var att rekrytera försökspersoner (fp) från SOL. Försökspersonen placerades i experimentrummet framför den stora skärmen med Psychopy-programmet i gång och försågs med hörlurar och mikrofon. De fick tid att läsa experimentbeskrivningen och ombads att säga "Börja" när de var redo att starta. Under denna tid satt försöksledaren i ett annat rum med visuell tillgång till försökspersonen och styrde experimentet från utrustningen därifrån. Samtliga experiment spelades in från start till slut, det vill säga från att de säger ”Börja” tills att de lämnar experimentrummet.

Efter varje experiment gick försöksledare tillbaka till experiment-rummet där fp befann sig, med inspelningen i RSE fortfarande i gång och utförde ett post-testintervju. Därefter förklarade försöksledare vad syftet med experimentet var, sedan måste fp välja att ge sitt medgivande genom att skriva under ett informationsblad som förklarar experimentet samt hur vi kommer hantera deras inspelningsdata. De belönas med ett ICA-presentkort värt 50 kr för deras deltagande, och får skriva under även där att de mottagit presentkortet.

Första pilot

Denna pilotstudie syftade till att bedöma påverkan av auditiv feedback på korttidsminnet, med fokus på huruvida deltagarnas minne påverkades mer av vad de hörde sig själva säga jämfört med vad de faktiskt sa.

Efter att ha förberett RSE-systemet, kopplat mikrofonerna och hörlurarna korrekt och startat experimentet i psychopy, rekryterade jag en försöksperson (fp). Jag introducerade mig själv, förklarade experimentets syfte och procedur, och ledsagade försökspersonen till experimentrummet där hen placerades framför datorskärmen och fick sätta på sig headsetet. Därefter påbörjades experimentet.

Den exakta beskrivning av experimentet, vilket var det första som försökspersonen fick när de väl satte sig på den utsedda stolen framför skärmen:

Du kommer att få göra ett minnestest. I varje trial kommer du att få se tre enstaviga ord på skärmen, alla tre orden börjar på samma bokstav, till exempel ”p”, ”b”, eller ”a”. Du kommer att se orden under 10 sekunder, och då ska du memorera ordningen på orden. Sedan kommer ett fixeringskors att visas under 1 sekund, och sedan kommer en så kallad visuell metronom att starta. Denna består av först en grön cirkel, sen en röd cirkel, sen en grön cirkel igen. Varje cirkel visas i 1.5 sekunder, och din uppgift är då att i takt med denna visuella metronom säga de tre orden i rätt mening, i takt med metronomen, så t.ex., ”borg bock boll”. Direkt efter detta så visas ett frågetecken på skärmen, och då ska du återigen upprepa de tre orden i rätt ordning, men här behöver du inte följa någon metronom utan bara säger orden i din egen takt. Det är alltid samma tre ord som dyker upp på varje bokstav, men ordningen på orden är randomiserad varje gång de dyker upp, så du måste lära dig ordningen varje gång nya ord kommer upp på skärmen. Det är viktigt att du pratar ganska så monotont och inte ändrar din intonation över experimentets gång. Det är 48 st ord-sekvenser totalt, och testet tar ca 15 minuter.
När du är redo att börja, säg ”börja” i mikrofonen.

Lycka till!

Denna experimentella design syftade till att undersöka interaktionen mellan auditiv återkoppling och minneshämtning, särskilt om deltagarna mindes det de sa eller det de hörde. Under experimentets gång spelade jag in vissa förbestämda ord i specifika sekvenser. Dessa inspelningar varierade mellan att vara det första, andra eller tredje ordet i sekvensen. Vid ett senare tillfälle i experimentet sköt jag in dessa inspelade ord på samma platser i sekvensen som de ursprungligen spelades in. Till exempel, om jag spelade in ”borg” i sekvensen ”bock borg boll” och ”boll” i sekvensen ”boll bock borg”, skulle jag vid nästa sekvens ”borg boll bock” se till att deltagaren hörde ”boll borg bock”. Jag byter alltså ”borg” med ”boll” och vice versa. Försökspersonen är omedveten om att denna utbyten kommer utföras. Syftet var att undersöka vilken sekvens deltagaren upprepade vid frågetecknet när de skulle återge de tre orden i rätt ordning. Upprepar det de faktiskt säger eller det de hör sig själva säga?

Valet av tre-ordsekvenser motiverades av att det verkade vara det optimala antalet ord för denna typ av minnestest. Preliminära tester med fyra-ordsekvenser visade sig vara svåra att komma ihåg, medan två-ordsekvenser bedömdes vara för enkla.

Andra pilot

Tidigare forskning undersökte den inferentiella modellen genom användning av Stroop-test med syfte att undersöka hur auditiv feedback påverkar talproduktion och huruvida fps märkte avvikelser mellan vad de sa och vad de faktiskt hörde. I denna pilotstudie avsåg vi att undersöka om fynden från de tidigare experiment kunde generaliseras till andra experimentella upplägg med högre ekologisk validitet. Ekologisk validitet avser hur väl resultaten från ett experiment kan överföras till verkliga livssituationer. Genom att använda en konversationsupplägg där deltagarna diskuterar olika bilder, strävade vi efter att skapa en mer naturlig och realistisk talmiljö jämfört med det mer strukturerade Stroop-testet.

Jag fick då agera som en konfedererad deltagare, vilket innebär att jag låtsades vara en rekryterad försöksperson för att undersöka den inferentiella modellen i ett konversationssammanhang. Den exakta instruktionen som jag och den riktiga fp fick var som följer:

Ni kommer att få samtala fritt kring bilder. Ni kommer turas om att välja en bild ni vill prata om och ni kan säga vad ni vill som kommer upp i era tankar. Det måste inte ha specifikt med bilden att göra utan ni kan associera fritt. Ni kan prata så länge ni vill om varje bild. På skärmen visas kontinuerligt en visuell metronom i form av en stor cirkel som byter mellan två färger i en specifik rytm. Rytmen är samma under hela experimentet. Ni kommer att prata i takt med denna metronom genom att säga ett ord för varje färgbyte. När ni känner att ni inte har mer att säga kring bilden så byter ni bild och gör samma sak igen. Ni kommer att konversera i ungefär 20 min på det här sättet.

När ni är redo att börja så kan en av er börja med att välja en bild så sätter ni igång. Lycka till!

Användningen av metronomen syftade till att segmentera talet i enskilda ord. Detta möjliggjorde att försöksledaren kunde få en bättre inspelning av ett isolerat ord, för att sedan kunna utbyta det mot ett annat liknande ord i en lämplig kontext. Försökspersonen var även i denna studie inte medveten om att det skulle ske utbyten.

Tredje pilot

Denna tredje pilotstudie syftade till att undersöka hur snabbt talare kan avbryta sitt tal baserat på auditiv återkoppling. Enligt litteraturen skulle sådana snabba responser typiskt antyda involvering av en intern monitoreringskanal baserad på enbart auditiv feedback. Att visa att talare kan avbryta sig själva baserat på enbart extern feedback skulle utmana behovet av att postulera en intern monitoreringskanal för sådana avbrott.

Den exakta instruktionen som gavs till deltagarna var som följer:

"Du kommer att få tala fritt kring bilder. Du väljer själv en bild du vill prata om, och du kan säga vad du vill som kommer upp i dina tankar. Det måste inte ha specifikt med bilden att göra, utan du kan associera fritt. Du kan prata så länge du vill om varje bild. På skärmen visas kontinuerligt en visuell metronom, i form av en stor cirkel som byter mellan två färger i en specifik rytm. Rytmen är samma under hela experimentet. Du kommer att prata i takt med denna metronom genom att säga ett ord för varje färgbyte. När du känner att du inte har mer att säga kring bilden, så byter du bild och gör samma sak igen. Du kommer att prata i ungefär 20 minuter på det här sättet.

Under experimentets gång kommer din röst slumpmässigt att manipuleras på så sätt att du säger ett ord, men samtidigt hör dig själv säga ett annat ord. När detta händer ska du så snabbt du bara kan avbryta ditt tal. Det är viktigt att du slutar prata så snabbt du kan; vi är intresserade just av hur snabbt talare kan göra sådana avbrott. När du har slutat prata, kan du direkt fortsätta prata där du var innan manipuleringen och avbrottet skedde.

Innan själva experimentet börjar ska du få göra en testgenomgång genom att prata spontant under 30 sekunder vid tre olika hastigheter på metronomen. Välj ut en bild som du kan börja prata om när du testar metronomen. Du kommer sen få välja den hastighet du känner dig mest bekväm med att prata i.

När du är redo att börja, säg 'börja' i mikrofonen. Lycka till!"

Genom att instruera deltagarna att medvetet notera och reagera på diskrepanser mellan deras tal och hörselintryck, kunde vi samla in data som utmanar och potentiellt förfinar den befintliga förståelsen av interna och externa monitoreringskanaler i talproduktion. Vi märkte även i pilot 2 att det varierade i hur försökspersonen uppfattade takten på metronomen. Vi hade inte riktigt tänkt att folk varierar i hur snabbt de pratar, därför behövde vi ge fps i pilot 3 möjlighet att välja mellan 3 takt som de skulle prata kring (700ms, 800ms och 900ms). Detta belyser vikten av att utföra pilotstudier där man behöver hela tiden vara öppen att revidera sin experimentsupplägg.

4 Preliminära resultat

I den första piloten stötte vi på betydande svårigheter med timing av manipulationerna. Av 125 försök till manipulationer var endast 27 % korrekt timade, och endast 4 resulterade i att deltagarna upprepade vad de hörde – en nyckelindikator på en framgångsrik manipulation. Dessa problem fick oss att ompröva och så småningom pausa den första pilotstudie, eftersom den låga andelen framgångsrika manipulationer

ifrågasatte genomförbarheten av experimentupplägget under nuvarande förhållanden.

Dataanalysen för de övriga pilotstudierna pågår fortfarande, vilket innebär att jag för närvarande inte kan presentera några definitiva resultat.

5 Vad lärde du dig?

Teoretiska insikter

Genom vår forskning erhöll jag en djupare förståelse för den roll som auditiv feedback spelar i talproduktion, vilket går långt utöver enbart felmonitorering. Studien avslöjade att meningsskapande processer och känslan av handlingskraft i tal är dynamiska fenomen som inte enbart uppstår från förformade avsikter, utan från ett intrikat samspel mellan talarens språkliga processer och de kontextuella ledtrådarna omkring dem, inklusive deras egen auditiva feedback. Detta indikerar att vår agens och förståelse av vårt eget tal inte är helt transparent för oss; snarare formas och förhandlas de kontinuerligt genom interaktion med vår omgivning.

Min kritiska utvärdering av den inferentiella modellen för talproduktion är att även om den ger betydande insikter om hur mening konstrueras, kan den förbise distinktionen mellan medvetna och omedvetna aspekter av tal. Modellen hävdar att det kanske inte alltid finns en pre-artikulatorisk intention, men kan det inte tänkas finnas varierande medvetenhet kring talets olika dimensioner. Sant må det vara att man sällan är medveten om det exakta ordet man ska säga, men man kanske ändå är medveten om det kommer vara ett substantiv eller ett verb. Om inte så kan man vara medveten om kontexten, vilket då eliminerar många alternativ.

På liknande sätt som large language models (llms) som chat GPT som numera är populära, inte vet vad de ska säga, utan listar ut det med hjälp av statistiska beräkningar, så kan det vara så att vår hjärna gör något liknande. Det finns en statistisk relation mellan bokstäver i ett ord, mellan ord i mening, mellan meningar i ett stycke, osv. Och så begränsas allt detta ytterligare av den breda kontexten (till exempel samtalsämnet). När vi då gör ett utbyte, ju mer kontexten skiljer sig från det inskjutna och det faktiskt sagda ordet, desto mer sannolikt är det att försökspersonen kommer märka det.

Denna tolkning hjälper till att förklara några av de utmaningar vi stötte på, särskilt i konversationspilotstudien, där den visuella metronomens rytm möjligen gjorde försökspersonerna mer medvetna om sina talavsikter och därmed ökade sannolikheten för att upptäcka manipulationer.

Praktiska lärdomar

På en praktisk nivå underströk projektet komplexiteten och utmaningarna med att genomföra vetenskapliga experiment. Behovet av noggrann planering och disciplinerat utförande blev uppenbart, särskilt när man satte upp experiment som involverade sofistikerad teknik som RSE. Pilotstudier är oerhört viktiga, de gör det möjligt för oss att identifiera och åtgärda potentiella problem innan vi går vidare till fullskalig forskning. Dessa tidiga försök underströk vikten av flexibilitet och beredskapen att anpassa experimentdesigner som svar på oväntade fynd eller tekniska svårigheter.

Att arbeta i en teammiljö gav också betydande lärandemöjligheter. Samarbete med erfarna forskare och andra assistenter möjliggjorde ett rikt utbyte av idéer och tekniker, vilket ökade allas förståelse och färdigheter. Denna kollaborativa lärmiljö var avgörande för att anpassa våra forskningsstrategier och i min personliga utveckling som forskare.

6 Vad kunde ha gjorts annorlunda?

Med tanke på att vårt projekt fortfarande är i pilotfasen är det något för tidigt att peka ut specifika förändringar som definitivt skulle förbättra resultaten. Vi lär oss kontinuerligt av varje pilotstudie, gör justeringar och förfinar våra metoder. Denna iterativa process är avgörande eftersom den tillåter oss att gradvis identifiera vad som fungerar och vad som inte fungerar inom ramen för vår experimentella design. Varje pilotstudie ger nya insikter som kontinuerligt formar vårt förhållningssätt och vår förståelse.

Potentiella förbättringar

En potentiell förbättring för framtida iterationer av liknande forskning, särskilt i miljöer som involverar konversationsdynamik, skulle kunna involvera integrering av avancerad teknologi som artificiell intelligens. Komplexiteten i att säkerställa korrekta utbyten – där verb byts ut mot verb, adjektiv mot adjektiv och så vidare – kräver oklanderlig timing och kontextuell lämplighet, vilket är utmanande att uppnå manuellt i realtid. Att implementera ett AI-system med maskininlärningsfunktioner skulle kunna dramatiskt förbättra effektiviteten och noggrannheten i talutbyten. Sådan teknik skulle kunna tränas för att hantera nyanserna av språklig kontext och timing, och automatiskt anpassa utbyten baserat på konversationsflödet. Detta skulle inte bara minska mänskliga fel utan skulle också möjliggöra en mer sömlös integrering av manipulationer, vilket skulle kunna leda till mer robust datainsamling.

Referenser

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22(1), 1-38.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say. Psychological Science, 25(6), 1198-1205.

Pickering, M. J., & Garrod, S. (2013). Forward Models and their Implications for Production, Comprehension, and Dialogue. Behavioral and Brain Sciences, 36(4), 377-392.