Institutionen för lingvistik
Stockholms Universitet

Audiovisuell perception av svenska vokaler med och utan motstridiga ledtrådar

Sammanfattning:

En bullrig miljö kan utgöra ett stort hinder att hänga med i ett samtal, i synnerhet vid nedsatt hörsel. I sådana situationer underlättar det om man kan se ansiktet på den man talar med. Visuell information om talarens ansikte bidrar till förståelsen av tal. När den visuella talsignalen inte stämmer överens med den auditiva kan illusioner uppstå. McGurk och MacDonald (1976) visade att ett auditivt [b] med ett visuellt [g] i de flesta fall uppfattades som [d]. Denna studie visar att McGurk-effekten även gäller vokaler. Audiovisuella stimuli med och utan motstridiga ledtrådar presenterades för lyssnare. Resultatet visar att läpprundningsgrad uppfattas genom synen snarare än hörseln. Käköppningsgrad uppfattas främst genom hörseln. Det visade sig att en i huvudsak manlig minoritet av lyssnarna inte litade till synintryck. Närvaro av visuell läpprundning uppmärksammas mer än frånvaro.

Kandidatuppsats i fonetik
Niklas Öhrström
Handledare: Hartmut Traunmüller
HT 2004

Innehållsförteckning

1. Introduktion

När normalhörande personer lyssnar använder de sig av hörseln för att avkoda talet. Det visuella bidraget till perceptionen ska dock inte negligeras. Det är sedan länge känt att personer med nedsatt hörsel använder sig av läppavläsning med visuella ledtrådar som kompensation. Men även för normalhörande utgör den visuella signalen en viktig, bidragande komponent för avkodning av tal, speciellt om de akustiska förutsättningarna är dåliga (Mártony, 1974).

Sumby och Pollack (1954) visade i en studie att normalhörande försökspersoners förståelse av uppläsa spondeeord i en kontext av vitt brus ökade om de förutom auditiv information även fick tillgång till visuell information om talaren. Det visuella bidraget till förståelsen ökade ju starkare störningsbruset var i förhållande till den akustiska talsignalen. Försökpersonerna hade inte utsatts för någon träning i läppavläsning. I en liknande studie har dessa resultat bekräftats av Erber (1969).

De ovannämnda resultaten visar att läppavläsning är något som normalhörande använder sig av i bullriga miljöer. Vad är då möjligt att läsa på läpparna? Amcoff (1970) undersökte svensktalande, normalhörande försökspersoners förmåga att läsa på läpparna. Syftet med studien var att dela in de svenska fonemen i visuellt kontrastiva klasser, ”visiofonem”, för att ge underlag till utformning av ett avläsestödssystem för hörselskadade. Studien visade att konsonanterna visuellt kunde delas in i (1) bilabialer, (2) rundade labialer, (3) labiodentaler samt (4) icke-labialer. De långa vokalerna kunde visuellt delas in i (1) orundade, (2) rundade utan protrusion samt (3) rundade med protrusion. Sammantaget tyder dessa resultat på att den visuella diskriminationsförmågan i hög grad baseras på labiala särdrag.

Forskningen kring audiovisuell perception tog ny fart vid upptäckten av McGurk-effekten (McGurk och MacDonald, 1976). En film med ett ansikte visades som uttalade nonsensordet, /gaga/. Den auditiva signalen hade bytts ut och ersatts med en annan auditiv signal, /baba/. Försökspersoner som hade sett filmen rapporterade att de hade hört /dada/, dvs. något som inte stämde överens med någon av de visuella eller auditiva signalerna. Ändrade man förhållandet så att det visuella stimulit /baba/ synkroniserades med det auditiva /gaga/, rapporterade en stor del av försökspersonerna att de hade hört /bagba/, dvs. en

kombination av de visuella och auditiva signalerna. Perception av tal hade fram till dess behandlats som en rent auditiv process (förutom när talsignalen störtes av brus). McGurk-effekten gav en antydan att talperception snarare var en bimodal process där stimuli från den auditiva respektive visuella modaliteten integreras till ett enhetligt percept (Risberg och Agelfors, 1978; Jordan och Sergeant, 2000; Colin et al., 2002; Green et al., 1991; Hietanen et al., 2001).

Upptäckten att perceptionen är en process där information från två olika modaliteter integreras har haft stor inverkan på forskningen kring talperception (Massaro, 1999). En central fråga är hur det integrerade perceptionsobjektet ska karakteriseras (Green, 1996): Består det av artikulatoriska gester eller är perceptionsobjektet till sin natur auditivt? Karakteristiskt för McGurk-fenomenet är att försökspersoner rapporterar att de har *hört* något som i själva verket inte bara är en produkt av den auditiva signalen utan även den visuella. Sams et al. (1991) gjorde neuromagnetiska mätningar på försökspersoners vänstra hjärnhalva när de utsattes för McGurk-stimuli. En viktig upptäckt de gjorde var att visuell information från läpprörelser påverkar aktiviteten i hörselbarken.

McGurk-effekten har undersöks under olika förhållanden. En slutsats som man dragit är att McGurk-effekten är robust. Om man synkroniserar ett manligt ansikte med en kvinnlig röst eller vice versa är McGurk-effekten tydlig i jämförelse med om ansiktet och rösten hade kommit från en och samma person (Green et al., 1991). Hietanen et al. (2001) undersökte McGurk-effekten med en manipulerad visuell signal. Manipulationen gick ut på att olika delar av ansiktet (ögon, näsa och mun) bytte plats med varandra. Så länge omkonfigurationen var symmetrisk var McGurk-effekten tydlig.

Vid audiovisuell integration är det inte nödvändigt att den visuella informationen om talarens ansikte är fullständig. Rosenblum och Saldaña (1996) genomförde en studie där bl.a. auditiva /ba/-stavelser hade synkroniserats med visuella /va/-stavelser. Dessa stimuli presenterades för försökspersoner under två olika betingelser: (1) Talarens ansikte var fullständigt upplyst, (2) Talarens ansikte representerades av upplysta punkter fastsatta på ansiktet. I övrigt var bilden svart. I den första betingelsen rapporterade i stort sätt samtliga försökspersoner att de hade hört /va/. Detta betydde att deras perception av artikulationsställe i stor utsträckning baserades på visuell information. Det mest

slående var att försökspersonerna rapporterade att de hade hört /va/ även då ansiktet endast representerades av upplysta punkter.

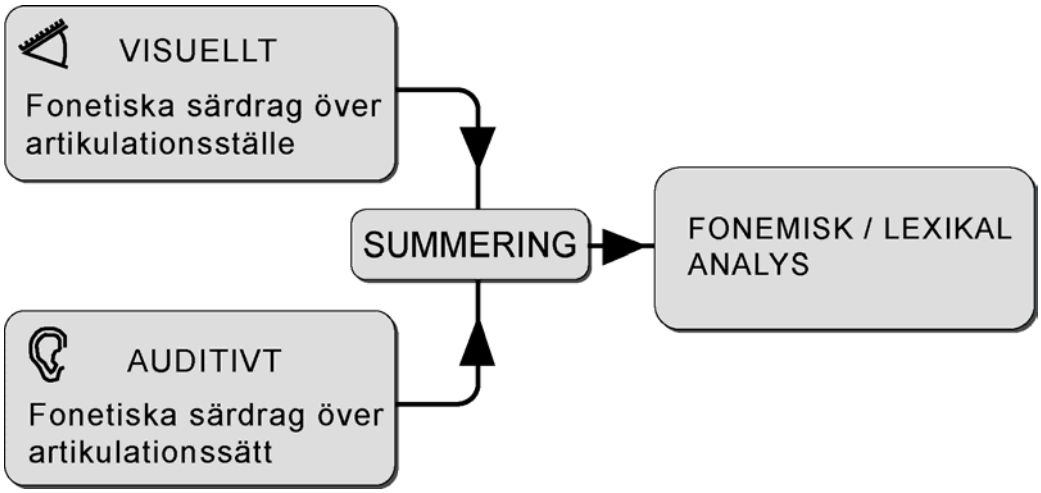
McGurk-effekten har undersökts under många olika betingelser och visat sig vara robust. Effekten har dock uppvisat avsevärda kulturella skillnader. Sekiyama och Tohkura (1993) visade i en tvärspråklig studie att McGurk-effekten hos modersmålstalare av japanska var mycket lägre än hos talare av amerikansk engelska. Resultatet tydde dock på att McGurk-effekten blev starkare om talaren var utlänning. Hayashi och Sekiyama (1998) genomförde en liknande jämförande studie mellan modersmålstalare av japanska och modersmålstalare av olika kinesiska dialekter. Båda grupperna uppvisade en låg McGurk-effekt då talaren talade samma språk som försökspersonerna. I fallet då talaren var utlänning blev effekten hos de japanska försökspersonerna förstärkt. Men hos de kinesiska försökspersonerna ökade inte effekten om talaren var utlänning.

Skillnader mellan kön i förmågan att läsa på läpparna har observerats. Johnson et al. (1988) visade i en studie, där modersmålstalare av engelska deltog, att kvinnor lyckas bättre än män på att känna igen fraser med endast visuell information från talarens ansikte. Den observerade skillnaden mellan könen torde även ge utslag på McGurk-test. Aloufy et al. (1996) visade att kvinnor i större utsträckning än män påverkas av visuella ledtrådar i McGurk-test. Men könsskillnaderna visade sig vara mycket mindre hos hebreisktalande än talare av amerikansk engelska.

Det finns flera studier som visar hur visuella ledtrådar bidrar till förståelsen av tal vid (1) hörselnedsättning, (2) när talsignalen störs av brus och (3) när de auditiva och visuella ledtrådarna är motstridiga (McGurk-effekt). Vid utformningen av modeller för audiovisuell perception måste man ta hänsyn till dessa fakta. Det har kommit en del förslag på modeller för audiovisuell perception. Summerfield (1987) presenterade olika förslag på modeller över audiovisuell integration.

I modeller som bygger på fonetiska särdrag integreras vissa särdrag från den visuella modaliteten och andra särdrag från den auditiva modaliteten. Ett exempel på denna sorts modeller är VPAM (Visual: Place, Acoustical: Manner). Denna modell bygger på hypotesen att den visuella och auditiva informationen kategoriseras i särdrag innan själva integrationen äger rum. Integrationen är en

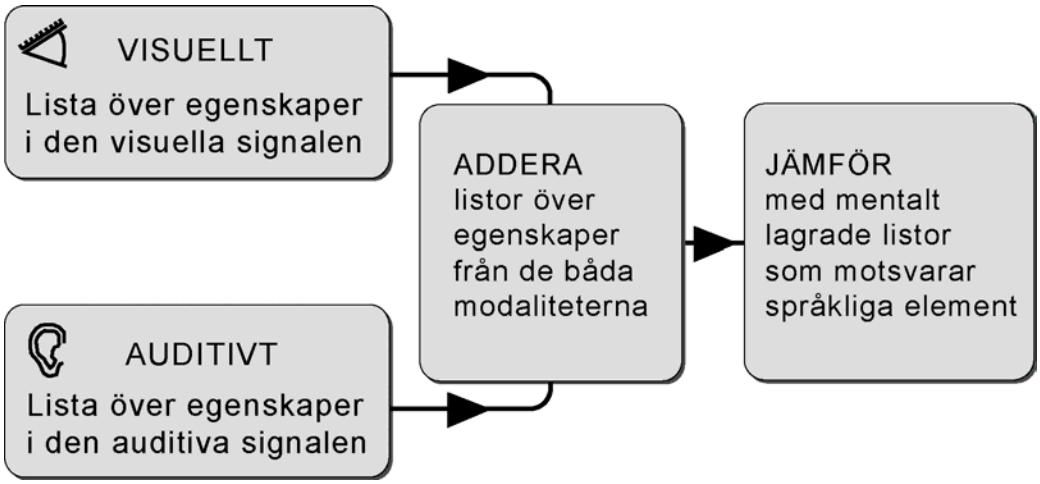
selektiv process där särdrag som rör artikulationsställe kommer från den visuella modaliteten och där särdrag som rör artikulationssätt kommer från den auditiva modaliteten (se figur 1). Enligt denna modell skulle ett visuellt [sɒl] kombinerat med ett auditivt [kɒl] percipieras som [tɒl]. Denna modell skulle dock inte kunna predicera McGurk-effekten eftersom ett visuellt [gaga] synkroniserat med ett auditivt [baba], enligt modellen skulle percipieras som [gaga] (enligt McGurk och MacDonald (1976) kommer [dada] att percipieras).



Figur 1. Audiovisuell perception av konsonanter enligt VPAM-hypotesen. Särdrag som rör artikulationsställe extraheras från den visuella modaliteten medan särdrag som rör artikulationssätt extraheras från den auditiva modaliteten.

Det finns även modeller för audiovisuell integration där informationen från den visuella modaliteten adderas med informationen från den auditiva modaliteten.

Informationen från dessa observationer matchas sedan mot lagrad information över de språkliga elementen (se figur 2).



Figur 2. I denna modell över audiovisuell perception erhålls en lista med egenskaper från den visuella respektive auditiva signalen. Dessa listor adderas och jämförs med lagrade listor (hos lyssnaren) som motsvarar fonem, stavelser eller ord.

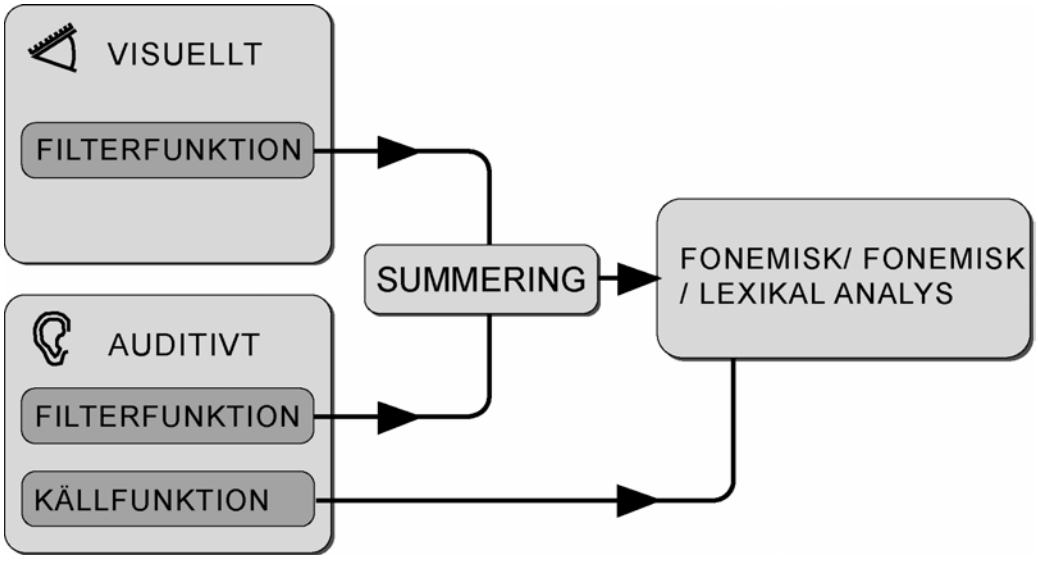
En variant av denna modell är ”the fuzzy logical model of perception” (FLMP) presenterad av Massaro (1987). I denna modell låter man en given akustisk eller optisk signal relateras till lyssnarens lagrade språkliga element. Man säger att en given signal motsvarar prototypen av ett visst lagrat språkligt element till en viss grad t (x), där x är en händelse där den akustiska (eller visuella) signalen identifieras med en prototyp av ett visst språkligt segment. t(x) = 0 betyder att signalen inte kan identifieras med segmentet och t(x) = 1 betyder att signalen fullt ut kan identifieras med segmentet. Funktionen t (x) kan således anta värden mellan 0 och 1. För att ta reda på till vilken grad en audiovisuell signal kan kopplas till ett visst segment tar man händelsen, x_A (en given auditiv signal kan kopplas till ett prototypiskt segment *a*) och x_V, (en given visuell signal kan kopplas till ett prototypiskt segment *a*). Den audiovisuella signalen kan då identifieras med segmentet *a* till en grad av $t(x_A \Lambda x_V) \cdot$ *normeringsfaktor*. Den

logiska operationen, Λ (disjunktion), mellan två händelser är definierad enligt ekvation 1.

$$t(x_A \Lambda x_V) = t(x_A) \cdot t(x_V) \qquad (1)$$

Ett problem med att ta disjunktionen mellan det auditiva och visuella fallet är att $t(x_A) \cdot t(x_V) \leq t(x_A)$. Detta betyder att även om de auditiva och visuella ledtrådarna inte är motstridiga skulle den visuella signalen försämra perceptionsförmågan. För att kompensera för att den logiska operationen sänker värdet på t(x) måste man multiplicera värdet med en normeringsfaktor.

Källa-filter-modellen för talproduktion bygger på att den akustiska talvågen är produkten av ett källspektrum (F_0 och dess multipler) och en filterfunktion (beroende av ansatsrörets konfiguration). En källa-filter-modell för talperception presenterades av Summerfield (1987). Enligt källa-filter-modellen för audiovisuell talperception demoduleras talvågen i källfunktion och i filterfunktion. Filterfunktionen integreras med den visuella informationen mottagaren har om talarens munrörelser. Den modulerade filterfunktionen och källfunktionen medverkar sedan till kategorisering av bl.a. de fonetiska elementen (se figur 3).



Figur 3. I källa-filter-teorin för talperception demoduleras talsignalen i en källfunktion och en filterfunktion. Informationen från den visuella signalen modifierar den auditiva filterfunktionen. Den modifierade filterfunktionen och källfunktionen bidrar till kategorisering.

McGurk-effekten hos klusiler har observerats och bekräftats genom ett stort antal studier. Man borde då fråga sig om inte samma fenomen skulle kunna uppträda hos vokaler. Summerfield och McGrath (1984) genomförde en studie där serier av auditiva vokaler i [bVd]-ram presenterades synkront med visuella ansikten som uttalade [bud], [bid] och [bɑd]. Resultatet från den studien visade att fonemgränserna i den auditiva vokalrymden för den auditivt presenterade vokalen förflyttades en liten bit i riktning mot den visuellt presenterade vokalen.

En slutsats som dragits från McGurk-experiment med klusiler är att närvaro eller frånvaro av labialisering är viktig visuell ledtråd för den perceptuella integrationen. Vid studier av läppavläsning har man funnit labiala särdrag är en viktig nyckel för att visuellt kunna skilja på olika konsonanter. Man har funnit att detta även gäller vokaler (Amcoff, 1970; Traunmüller, 1979). Denna studies syfte är att undersöka den audiovisuella integrationen hos vokaler i ett språk där läpprundning är ett oberoende distinktivt särdrag. Hypotesen är att läpprundning visuellt har stor påverkan på perceptionen av vokaler.

2. Metod

Denna studie har genomförts i tre faser. Till att börja med spelades fyra talare in på video, när de uttalade stavelser. Sedan manipulerades videoinspelningarna så att var och en av de visuella delarna av de inspelade stavelserna synkroniserades med var och en av de auditiva. Varje auditivt stimulus och varje visuellt stimulus presenterades var för sig. Slutligen fick lyssnare se och lyssna på det manipulerade talmaterialet och avgöra vilken vokal de hade hört.

2.1 Försökspersoner

2.1.1 Talare

Talarna var två män (på 29 och 45 år) och två kvinnor (på 21 och 29 år). Alla var verksamma inom Institutionen för lingvistik på Stockholms universitet. Alla var modersmålstalare av svenska. Ingen bedömdes inte ha någon utpräglad dialekt. Ingen ekonomisk ersättning utgick till talarna.

2.1.2 Lyssnare

I perceptionstestet deltog 21 lyssnare. Av dessa var 11 kvinnor och 10 män. Alla var modersmålstalare av svenska, hade normal hörsel samt normal syn eller korrigerad syn. Ingen av försökspersonerna hade varit aktiv eller verksam inom ämnesområdena lingvistik eller fonetik. Medelåldern var 26,6 år och åldersspannet var 16 år till 49 år. Tolv var studerande, åtta var yrkesverksamma och en var arbetssökande. Ingen ekonomisk ersättning utgick till lyssnarna.

2.2 Talmaterial

2.2.1 Val av talmaterial

I undersökningen avsågs att undersöka hur perception av svenskans långa vokaler kan påverkas när de visuella och auditiva ledtrådarna är motstridiga. För

att perceptionstestet skulle tjäna sitt syfte, togs vid utformningen hänsyn till följande faktorer:

1. De undersökta vokalerna måste kunna delas in i oberoende distinktiva särdrag som är visuellt och auditivt möjliga att uppfatta. I svenskan uppfyller särdragen käköppningsgrad samt läpprundning dessa krav.
2. Kontextens påverkan på vokalernas produktion med avseende på läpprundning och käköppningsgrad ska minimeras. Därför bör man undvika en labial kontext.
3. Om stimuli lätt kan förväxlas med något ord i det svenska ordförrådet finns det stor risk att lyssnare i stor utsträckning väljer ett sådant ord i ett perceptionstest (Amcoff, 1970). Därför ska förväxlingsbara ord vara nonsensord.

För att uppfylla dessa faktorer valdes att undersöka de svenska långa vokalerna; /i/, /y/, /e/ och /ø/ i en [g_g]-kontext.

2.2.2 Inspelningsprocedur

Fyra talare (se kapitel 2.1.1) spelades in digitalt på video. Inspelningarna ägde rum i ett perceptionslaboratorium på Institutionen för lingvistik, Stockholms universitet. Rummet var upplyst så att inga skuggkonturer syntes på talarnas ansikten. Vid inspelningen filmades talarnas ansikten med en videokamera på 2,5 meters avstånd. Kameran var på samma höjd som talarnas ansikten. Talarnas ansikten zoomades in så att bildens nedre kant låg i höjd med struphuvudet och bildens övre kant låg 2 cm ovanför talarnas huvud. Ljudet spelades in med en mikrofon på 50 centimeters horisontellt avstånd från talarens mun. Mikrofonen var riktad mot talarens haka och vinkeln i förhållande till munnen var 45 grader. Som bakgrund användes ett ljusblått skynke. Videokameran var en Panasonic NV-DS11. Vid inspelningen användes en extern kondensatormikrofon; AKG CK 93, förstärkare; AKG SE 300B och batterihållare; AKG B18.

Vid videoinspelningen uttalade talarna stavelserna /gig/, /gyg/, /geg/ och /gøg/. Stor vikt lades vid att stavelsernas durationer i största möjliga mån skulle vara lika långa. Talarna instruerades även att uttala stavelserna på ett naturligt sätt utan att ge avkall på tydlig artikulation.

2.2.3 Manipulation av inspelat material

Till perceptionstestet manipulerades det inspelade materialet till audiovisuella stimuli med (1) överensstämmande ledtrådar från de visuella och auditiva modaliteterna, (2) motstridiga ledtrådar från de visuella och auditiva modaliteterna. Dessa stimuli skapades genom att synkronisera vart och ett av de auditiva stavelserna med vart och ett av de visuella stavelserna. I perceptionsstestet förekom även unimodala stimuli med (3) endast auditiva ledtrådar och (4) endast visuella ledtrådar. Redigeringen skedde med hjälp av Adobe Premiere 6.0. På så sätt erhölls 24 stimuli från var och en av de fyra talarna (se tabell 1). I perceptionstestet förekom varje stimulus två gånger i randomiserad ordning. Följaktligen bestod perceptionstestet av totalt 192 stimuli. Perceptionstestet var uppdelat på 24 block med 8 stimuli i varje. Stimuli presenterades med fyra sekunders intervall och pausen mellan blocken varade i nio sekunder. Testet spelades upp med hjälp av Windows Media Player. Ett mål med redigeringen var att synkronisera explosionerna hos de stavelseinitiala [g]:na. Tidsdiskrepansen mellan det auditivt och visuellt presenterade stimulus underskred alltid 5 ms.

Tabell 1. De 24 olika stimuli som erhölls av var och en de fyra talarna.

Auditivt	Visuellt
/gig/	/gig/
/gig/	/gyg/
/gig/	/geg/
/gig/	/gøg/
/gyg/	/gig/
/gyg/	/gyg/
/gyg/	/geg/
/gyg/	/gøg/

Auditivt	Visuellt
/geg/	/gig/
/geg/	/gyg/
/geg/	/geg/
/geg/	/gøg/
/gøg/	/gig/
/gøg/	/gyg/
/gøg/	/geg/
/gøg/	/gøg/

Auditivt	Visuellt
*	/gig/
*	/gyg/
*	/geg/
*	/gøg/
/gig/	*
/gyg/	*
/geg/	*
/gøg/	*

2.3 Perceptionstest

I perceptionstestet deltog lyssnarna en och en. De satt med ansiktena 60 cm från en dataskärm. Talarnas ansikten som uppträdde på dataskärmen var ca 17 cm höga. Lyssnarna bar hörlurar av märket AKG K135. De instruerades muntligen och skriftligen att fokusera blicken på den presenterade talarens ansikte och skriva upp på ett svarsformulär vilken vokal de hört/uppfattat. Lyssnarna förbereddes även på att vissa stimuli skulle vara enbart visuella eller auditiva. I det förra fallet instruerades lyssnarna att försöka läsa på läpparna vilken vokal som uttalats. I det senare fallet (enbart auditivt stimuli) instruerades de att skriva upp vilken vokal de hört. Lyssnarna fick välja mellan de nio svenska bokstäver som representerar vokaler. Endast ett svar var tillåtet.

Under perceptionstestet övervakades lyssnarna för att kontrollera att de var koncentrerade och hela tiden höll blicken vänd mot talarens ansikte. Innan själva perceptionstestet ägde rum fick lyssnarna öva sig på åtta stimuli för att bekanta sig med testet. Hela sessionen varade i ungefär 20 min.

3. Analys och resultat

Resultaten från perceptionstestet visade preliminärt att lyssnarnas perceptuella beteende varierade. Detta föranledde författaren att utröna om lyssnarna kunde delas in i olika kategorier på basis av deras svar. För var och en av lyssnarnas svar genomfördes en stegvis regressionsanalys för att undersöka hur stor del av variansen hos svaren som kunde förklaras med de oberoende faktorerna ”auditiv öppningsgrad”, ”auditiv rundningsgrad”, ”visuell öppningsgrad” och ”visuell rundningsgrad”. Analysen genomfördes genom att tilldela de olika fonemen numeriska värden (se tabell 2). Variansen mättes för de stimuli där den visuellt presenterade vokalen skiljde sig mot den auditivt presenterade vokalen med de båda särdragen rundningsgrad och öppningsgrad. Detta betyder att svaren på följande auditiva-visuella stimuli analyserades; /gig/-/gøg/, /gyg/-/geg/, /geg/-/gyg/ och /gøg/-/gig/. Till de statistiska beräkningarna användes programpaketet SPSS.

Tabell 2. De numeriska värden som varje fonem tilldelades.

Fonem	i	y	e	ø	ɛ	ɒ	o
Öppningsgrad	0	0	1	1	1	2	2
Rundningsgrad	0	1	0	1	1	0	1

Regressionsanalysen visade att ”auditiv öppningsgrad” förklarade det mesta av variansen hos samtliga lyssnares svar. För 16 av de 21 lyssnarna (Grupp 1) kom ”visuell rundningsgrad” på andra plats. För de resterande fem lyssnarna (Grupp 2) kom ”auditiv rundningsgrad” på andra plats. I den fortsatta analysen kommer dessa båda grupper att behandlas separat. Grupp 1 bestod av sex av de tio manliga lyssnarna och tio av de elva kvinnliga lyssnarna. Värt att notera är att Grupp 2 bestod av fyra av de tio manliga lyssnarna (40 %) men bara en av de elva kvinnliga (9 %).

3.1 Perception utan motstridiga ledtrådar

I kontrollsyfte användes stimuli utan motstridiga ledtrådar. Detta betydde att den audiovisuella signalen inte var manipulerad. Ett audiovisuellt presenterat [i] uppfattades av Grupp 1 som [y] i ett fall. I de resterande 127 fallen percipierades [i]. De andra vokalerna, [y], [e] och [ø] uppfattades korrekt till 100 %. I Grupp 2 förväxlades ett presenterat [i] som [y] i ett fall. I de resterande 39 fallen percipierades [i]. Grupp 2 uppfattade de andra vokalerna, [y], [e] och [ø] korrekt till 100 %.

3.2 Perception av endast en modalitet

När lyssnarna utsattes för de stimuli där endast de visuella ledtrådarna var närvarande, lyckades både Grupp 1 och Grupp 2 att uppfatta distinktionen mellan rundade och icke-rundade vokaler. Grupp 1 svarade korrekt i 97 % av fallen medan Grupp 2 svarade korrekt i 95 %. Lyssnarnas resultat vad gäller uppfattning av käköppningsgrad var lägre. Grupp 1 lyckades bäst och hade 73 % korrekt uppfattning av käköppningsgrad. Grupp 2 svarade rätt i 63 % av fallen. Resultaten över lyssnarnas svar med endast visuella ledtrådar är presenterade i tabell 3A och 3B.

Tabell 3A. Förväxlingsmatris för Grupp 1 (tio kvinnor och sex män). Raderna visar visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Visuellt	i	y	e	ø	ɛ	ɒ	o
i	58	2	34	5	1		
y		60		40			
e	16	1	80	1	2		
ø		8	1	90		1	1

Tabell 3B. Förväxlingsmatris för Grupp 2 (en kvinna och fyra män). Raderna visar visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Visuellt	i	y	e	ø	ɛ	ɒ	o
i	35	3	60	3			
y	3	63		35			
e	15	8	78				
ø		25	3	67		5	

Vid perception av stimuli med endast auditiva ledtrådar lyckades båda grupperna uppfatta distinktionen i käköppningsgrad. Grupp 1 hade 98 % korrekt uppfattad käköppningsgrad och Grupp 2 hade 100 % rätt uppfattad käköppningsgrad. Läpprundningen uppfattades korrekt i 98 % för Grupp 1 och 100 % för Grupp 2. Auditivt uppfattades läpprundning korrekt till 92 % av Grupp 1 och 97 % av Grupp 2. Resultaten över svaren på stimuli med endast auditiva ledtrådar återfinns i tabell 4A och 4B.

Tabell 4A. Förväxlingsmatris för Grupp 1 (tio kvinnor och sex män). Raderna visar auditivt presenterade vokaler och kolumner visar percipierade vokaler (%).

Auditivt	i	y	e	ø	ɛ	ɒ	o
i	91	9					
y	3	97					
e			84	16			
ø				95	5		

Tabell 4B. Förväxlingsmatris för Grupp 2 (en kvinna och fyra män). Raderna visar auditivt presenterade vokaler och kolumner visar percipierade vokaler (%).

Auditivt	i	y	e	ø	ɛ	ɒ	o
i	93	8					
y	5	95					
e			100				
ø				100			

3.3 Perception med motstridiga ledtrådar

3.3.1 Motstridighet med ett särdrag - Läpprundning

När stimuli, där den visuella signalen signalerar läpprundning medan den auditiva signalerar läppspridning och vice versa, presenteras skiljer sig de båda gruppernas resultat markant. Grupp 1 uppfattade läpprundning genom den visuella signalen i 78 % av fallen och genom den auditiva signalen i 22 %. Grupp 2 uppfattade läpprundning genom den visuella signalen i 16 % av fallen och genom den auditiva signalen i 84 % av fallen. De båda gruppernas resultat från dessa stimuli återfinns på tabell 5A och 5B.

Tabell 5A. Förväxlingsmatris för Grupp 1 (tio kvinnor och sex män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	y	5	94		1			
y	i	77	22	1				
e	ø			15	85			
ø	e			6	46	48		

Tabell 5B. Förväxlingsmatris för Grupp 2 (en kvinna och fyra män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	y	78	23					
y	i	20	80					
e	ø			78	23			
ø	e				85	15		

3.3.2 Motstridighet med ett särdrag - Käköppningsgrad

När den visuella signalen skiljde sig från den auditiva med särdraget käköppningsgrad, litade båda grupperna på hörseln till 100 % medan synen inte

gav något bidrag till uppfattningen av detta särdrag. De båda gruppernas resultat från dessa stimuli återfinns på tabell 6A och 6B.

Tabell 6A. Förväxlingsmatris för Grupp 1 (tio kvinnor och sex män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	e	100						
y	ø		100					
e	i			99	1			
ø	y				100			

Tabell 6B. Förväxlingsmatris för Grupp 2 (en kvinna och fyra män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	e	98	3					
y	ø		100					
e	i			100				
ø	y				100			

3.3.3 Motstridighet med två särdrag – Läpprundning och käköppningsgrad

Perceptionen av de stimuli, där den visuella signalen skiljde sig från den auditiva med de båda särdragen läpprundning och käköppningsgrad, kunde delas upp i tre kategorier: (1) Perception av endast visuella ledtrådar, (2) perception av endast auditiva ledtrådar och (3) perception där ledtrådar från både den visuella och auditiva modaliteten har sammansmält. I detta fall utgjordes samtliga sammansmältningar av uppfattad visuell rundningsgrad kombinerat med auditivt uppfattad käköppningsgrad. Resultaten från de två grupperna skiljde sig avsevärt. De två gruppernas svarsfördelning presenteras i tabell 7.

Tabell 7. Svarsfördelningen hos de båda grupperna vid motstridighet med två särdrag (%). Sammansmältning definieras som visuell rundning och auditiv öppningsgrad.

	Grupp 1	Grupp 2
Auditivt	22	74
Visuellt	2	0
Sammansmältn.	76	26

I de fall då ett visuellt /gig/ kombinerades med ett auditivt /gøg/ rapporterade lyssnarna i många fall att de uppfattat ett /gɛg/. Trots att /ø/ och /ɛ/ skiljer sig åt i fråga om käköppningsgrad har dessa svar fallit under kategorin ”sammansmältningar”. De båda gruppernas resultat från stimuli med motstridiga ledtrådar med två särdrag återfinns på tabell 8A och 8B.

Tabell 8A. Förväxlingsmatris för Grupp 1 (tio kvinnor och sex män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	ø	13	84		3			
y	e	84	13	3				
e	y			1	99			
ø	i			5	62	34		

Tabell 8B. Förväxlingsmatris för Grupp 2 (en kvinna och fyra män). Raderna visar auditivt och visuellt presenterade vokaler och kolumner visar percipierade vokaler (%).

Aud	Vis	i	y	e	ø	ɛ	ɒ	o
i	ø	70	30					
y	e	35	65					
e	y			65	35			
ø	i				98	3		

3.4 Asymmetri i visuell perception av läpprundning

Inom Grupp 1 var förväxlingsmönstret i rundningsgrad asymmetriskt. En vokal identifierades sällan som orundad när den visuellt presenterade vokalen var rundad, men när läpparna var synligt orundade var inte förväxlingar lika ovanliga (se tabell 9).

Tabell 9. Förväxlingsmatris över visuellt percipierad rundningsgrad. ”0”= visuellt orundad, ”1”= visuellt rundad. Raderna visar visuellt presenterad rundningsgrad och kolumner visar percipierad rundningsgrad (%).

	Grupp 1		Grupp2	
Rundning	0	1	0	1
0	85	15	66	35
1	3	97	30	71

3.5 Regressionsanalys

En stegvis regressionsanalys genomfördes för att utröna vilka faktorer som bidrog till perceptionen av rundningsgrad samt käköppningsgrad för Grupp 1 och Grupp 2. Varje vokal som förekom bland svaren tilldelades ett numeriskt värde för de två parametrarna rundningsgrad samt öppningsgrad (se tabell 2). De faktorer som undersöktes var *auditiv rundningsgrad, auditiv öppningsgrad, visuell rundningsgrad* samt *visuell öppningsgrad*. Analysen tog även hänsyn till interaktionsfaktorer som *auditiv öppningsgrad * auditiv rundningsgrad* och *visuell öppningsgrad * visuell rundningsgrad*. Resultaten från regressionsanalysen återfinns på tabell 10. För de båda grupperna bidrog de auditiva ledtrådarna starkt till perceptionen av öppningsgrad. När det gäller rundningsgrad skiljer sig de båda grupperna åt. Grupp 1 uppfattar rundningsgrad i första hand genom visuella ledtrådar. Värt att notera är att de auditiva ledtrådarna till läpprundning inte ens är signifikanta. Grupp 2 litade i första hand på hörseln vid perception av rundningsgrad. Visuella ledtrådar lämnade förvisso ett signifikant bidrag till perceptionen av rundningsgrad.

Tabell 10. Resultat från den stegvisa regressionsanalysen. De signifikanta faktorerna förklarade variansen till en grad (r^2). Ej signifikanta faktorer benämns e.s.

	Grupp 1		Grupp 2	
	rundn	öppn	rundn	öppn
r^2	0,926	0,965	0,972	0,995
Konstant	0,18	0,01	0,03	0,00
Aud öppn	e.s	0,99	e.s	1,02
Aud rund	e.s	e.s	0,77	e.s
Aud rund*öp	0,27	0,20	e.s	e.s
Vis öppn	e.s	e.s	e.s	e.s
Vis rund	0,77	e.s	0,22	e.s
Vis rund*öpp	e.s	e.s	e.s	e.s

4. Diskussion

4.1 McGurk-effekt för vokaler

Denna uppsats huvudfråga var att utröna hur vokaler med motstridiga ledtrådar perceptuellt integreras. Gäller McGurk-effekten även för vokaler?

Studien visade att ett auditivt [e] i synkroni med ett visuellt [y] i de flesta fall uppfattades som ett [ø]. När den auditiva vokalen, [i], presenterades i synkroni med den visuella vokalen, [ø], uppfattades det för det mesta som ett [y]. Ett visuellt [e] synkroniserat med ett auditivt [y] uppfattades oftast som [i]. Denna sorts percept kan ses som en variant av de sammansmältningar som uppträder då ett visuellt [gaga] presenterat i kombination med ett auditivt [baba] resulterar i ett percipierat [dada] (McGurk och MacDonald, 1976).

Som tidigare nämnts är McGurk-effekten inte symmetrisk: Då ett visuellt [baba] presenteras tillsammans med ett auditivt [gaga] sker ingen sammansmältning av de båda modaliteterna. I stället visade det sig att försökspersoner i stor utsträckning hade uppfattat en seriell kombination av de båda modaliteterna; [gaba], [gabga] eller [bagba] (McGurk och MacDonald, 1976). Colin et al. (2002) visade att seriella kombinationer är vanligare för tonlösa konsonanter än för tonande. En tolkning av dessa resultat skulle kunna vara att förekomsten av seriella kombinationer för vokaler skulle vara nästintill obefintlig. Denna studie har inte undersökt om fenomenet med seriella kombinationer uppträder hos audiovisuell perception av vokaler. En svårighet med att undersöka detta eventuella fenomen för vokaler är att vokaler, till skillnad från klusiler inte percipieras kategoriskt. (Strange, 1995). En seriell kombination av vokaler skulle uppfattas som en diftong. I många svenska dialekter motsvaras enskilda fonem av utpräglade diftonger (Elert, 2000). I denna studie rapporterade ett par av lyssnarna spontant (efter sessionen) att de hade hört diftonger som inte motsvarade talarnas dialekter (t.ex. skånska diftonger). Det är svårt att dra några slutsatser om förekomsten av seriella kombinationer på basis av vad några lyssnare spontant har rapporterat. Men man bör heller inte utesluta möjligheten att det skulle kunna föreligga en

asymmetri i perceptionen av vokaler analogt med vad som upptäcktes med klusiler av McGurk och MacDonald (1976).

4.2 Perceptuell integration

Vid audiovisuell integration förefaller det som att vissa särdrag percipieras från den visuella modaliteten och andra särdrag från den auditiva modaliteten. I detta fall har vi sett att käköppningsgrad i huvudsak percipieras genom auditiva ledtrådar. Detta gällde de båda grupperna trots att Grupp 1 även visade på god förmåga att uppfatta käköppningsgrad genom läppavläsning. En viktig upptäckt var att läpprundning hos Grupp 1 i princip endast uppfattas genom visuella ledtrådar. Detta är i linje med hypotesen att läpprundning visuellt har stor påverkan på perceptionen av vokaler. De auditiva ledtrådarna lämnade i sig inte ens något signifikant bidrag till den audiovisuella integrationen. Med tanke på att Grupp 1 lyckades att auditivt uppfatta läpprundning korrekt i hela 92 % av fallen är dessa resultat anmärkningsvärda. Att labialitet är prominent i den visuella modaliteten är förvisso ingen nyhet: McGurk och MacDonald (1976) visade att närvaron av labialitet är en viktig visuell ledtråd vid audiovisuell perception av klusiler. Denna studie har visat att detta inte bara gäller klusiler utan även vokaler.

Resultaten från denna undersökning tyder på att de flesta percipierar visuell rundningsgrad och auditiv käköppningsgrad. Ett undantag från detta mönster har tydligt visat sig: Ett auditivt [ø] kombinerat med ett visuellt [e] eller [i] uppfattades av Grupp 1 till stor del som [ɛ] och inte som ett [e]. Detta återspeglas i regressionsanalysens interaktionsfaktor (auditiv öppningsgrad*läpprundning) som lämnade ett signifikant bidrag till variansen i käköppningsgrad samt rundningsgrad. Att ett auditivt [ø] förväxlas med ett [ɛ] och inte [e] kan bero på akustikperceptuella faktorer: Både F_1 och F_2 är starka akustiska perceptuella ledtrådar. Eklund och Traunmüller (1997) visade att F_1 för det svenska uttalet av /ø/ ligger mittemellan /e/ och /ɛ/. F_2 för uttalet av /ø/ ligger mycket nära /ɛ/ men långt under /e/. Dessa resultat som visat att /ø/ akustiskt ligger närmare /ɛ/ än /e/ har nyligen bekräftats av Nilsson (2003) som genom formantmätningar av svenska vokaler presenterade resultat som kan

tolkas som att uttalet av /ø/ akustiskt ligger närmare /ɛ/ än /e/ (i en tvådimensionell F_1-F_2-rymd) och att detta i synnerhet gäller yngre talare.

Perceptionen av labialisering uppvisade ett asymmetriskt mönster hos Grupp 1: Det förekom att lyssnarna percipierade en rundad vokal när de visuella ledtrådarna signalerade frånvaro av läpprundning. Men när den visuellt presenterade vokalen var rundad percipierades mycket sällan en orundad vokal. Vad detta beror på kan man bara spekulera i, men det verkar som att lyssnare anser att orundade läppar till (skillnad från rundade) är omarkerat och inte uppmärksammas på samma sätt som synligt rundade läppar.

4.3 Skillnader mellan kön

Tidigare studier har visat att McGurk-effekten är mer framträdande hos kvinnor än hos män (Aloufy et al., 1996). Man har även visat att kvinnor generellt är bättre än män på läppavläsning (Johnson et al., 1988). Resultaten från den här studien går i samma riktning: En minoritet som bestod av 40 % av de manliga lyssnarna samt bara 9 % av de kvinnliga påverkades i ringa utsträckning av visuella ledtrådar vid perception av vokaler. Majoriteten som bestod av 60 % av de manliga lyssnarna samt hela 91 % av de kvinnliga påverkades i stor utsträckning av visuella ledtrådar vid perception av vokaler.

Förklaringen till att visuella ledtrådar spelar en större roll hos kvinnors perception än mäns perception sägs ligga i att kvinnor använder sig av annorlunda strategier vid talperception: De är mer benägna att titta på talarens ansikte. Förutom könsskillnader är det värt att tillägga att skillnader mellan kulturer föreligger. Som tidigare nämnts är talperceptionen hos japansktalande inte påverkad av visuella ledtrådar så länge talaren inte är utlänning. Detta sägs bero på att det i den japanska kulturen är oartigt att stirra någon i ansiktet. (Sekiyama och Tohkura, 1993; Hayashi och Sekiyama, 1998).

4.4 Modeller

I början av denna uppsats nämndes ett antal modeller över audiovisuell talperception. Hur kompatibla är dessa med de upptäckter som redovisats gällande audiovisuell perception av vokaler?

En särdragsbaserad modell skulle kunna beskriva audiovisuell integration av vokaler ganska bra (så som den beskrivs i denna uppsats). Ett förslag skulle vara att man percipierar läpprundning genom den visuella modaliteten och käköppningsgrad genom den auditiva. En sådan modell skulle dock inte kunna beskriva att ett auditivt [ø] kombinerat med ett visuellt [i] eller [e] percipieras som ett [ɛ]. En sådan modell skulle i detta fall predicera ett [e]. Det verkar som att det produktionsbaserade särdraget käköppningsgrad i denna modell borde ersättas av ett akustiskt särdrag. Men att blanda ihop produktionsbaserade särdrag med akustiska särdrag i en och samma modell kan te sig lite klumpigt. Ett alternativ för att kunna kringgå problemet kring särdraget käköppningsgrad är att komplettera modellen med en *ad hoc*-lösning. Men man kan fråga sig hur relevant en modell är om den fordrar *ad hoc*.

The fuzzy logical model of perception (FLMP) (Massaro, 1987) kan på ett adekvat sätt predicera McGurk-effekten för klusiler. Ett problem med FLMP vid perception av vokaler är att de inkommande språkliga signalerna matchas mot lagrade prototyper av språkliga segment. Hur ska man kunna definiera en prototyp av en vokal då fonemgränserna i vokalrymden inte är distinkta (Strange, 1995)?

Enligt källa-filter-teorin för perception separeras den inkommande talsignalen i en källfunktion och en filterfunktion. Vid audiovisuell perception modifieras filterfunktionen av den visuella signalen. Den modifierade filterfunktionen och källsignalen medverkar till kategorisering av de fonetiska elementen. Denna modell är lite oprecist formulerad men skulle kunna predicera McGurk-effekten och även de upptäckter som beskrivits i denna uppsats.

Medan källa-filter-teorin fungerar bra för talproduktion kan man fråga sig hur relevant den är för talperception. Hur kan en lyssnare separera ut en källfunktion och en filterfunktion ur den akustiska talsignalen? Visserligen är det ingen svårighet för en lyssnare att perceptuellt sortera ut en akustisk talsignal ur omgivande buller. I detta fall kan lyssnaren sortera ut talsignalen eftersom hon/han vet hur en talsignal låter. I fallet med källsignalen kan lyssnaren omöjligen veta hur den låter omodulerad. Därför kan man också ifrågasätta hur lyssnaren perceptuellt ska kunna sortera ut den ur talsignalen. Modulationsteorin (Traunmüller, 1994) löser detta problem genom att beskriva talsignalen som en ickemodulerad talsignal (neutral vokal, schwa) som

modulerats av språkliga gester. Den ickemodulerade signalen benämns som bärsignal och innehåller information om talarens ålder, kön och emotionellt tillstånd samt var talaren befinner sig. Modulationssignalen innehåller rent språklig information. Vid talperception demoduleras talsignalen i bärsignal och modulationssignal. Modulationsteorin är inte riktigt komplett vad gäller audiovisuell perception så det skulle vara intressant att se hur upptäckterna som beskrivits i denna uppsats kan bidra till en komplettering av modulationsteorin.

5. Slutsatser

Studien visar att lyssnare, under betingelsen att de auditiva och visuella ledtrådarna är motstridiga, tenderar att uppfatta rundningsgrad genom visuella ledtrådar och käköppningsgrad genom auditiva. Att ett auditivt [ø] kombinerat med ett visuellt [i] eller [e] snarare uppfattas som [ɛ] än som [e] förklaras med att uttalet av /ø/ auditivt ligger närmare /ɛ/ än /e/. Dessa resultat kan ses som en variant på de sammansmältningar som upptäcktes av McGurk och MacDonald (1976). Lyssnarnas svar skiljde sig åt och man kunde dra slutsatsen att kvinnor i stor utsträckning använder sig av visuella ledtrådar vid perception av vokaler med motstridiga ledtrådar. Män skiljer sig åt i sina svar men omkring hälften av dem använder sig i stor utsträckning av visuella ledtrådar vid audiovisuell perception av vokaler med motstridiga ledtrådar. Närvaro av läpprundning uppmärksammades mer än frånvaro.

Ingen av de presenterade modellerna över audiovisuell perception var riktigt kompatibel med de resultat som presenterats i denna uppsats. Det skulle dock vara intressant att se hur modulationsteorin skulle kunna behandla denna sorts fenomen.

6. Referenser

Aloufy S., Lapidot M., och Myslobodsky (1996) Differences in susceptibility to the ”blending illusion” among native Hebrew and English Speakers. Brain and Language 53, 51-57

Amcoff S. (1970) Visuell perception av talljud och avläsestöd för hörselskadade. Rapport Nr. 7, LSH Uppsala, Pedagogiska institutionen.

Colin C., Radeau M., Deltenre P., Demolin D. och Soquet A. (2002) The role of sound intensity and stop-consonant voicing on McGurk fusions and combinations. European Journal of Cognitive Psychology, 14(4), 475-491.

Eklund I. och Traunmüller H. (1997) Comparative study of male and female whispered and phonated versions of the long vowels of Swedish. Phonetica 54, 1-21

Elert C.-C. (2000) Allmän och svensk fonetik. Stockholm: Norstedts. Åttonde omarbetade upplagan.

Erber N.P. (1969) Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech and Hearing Research. 12. 423-425.

Green K.P., Kuhl P.K., Meltzoff A.N., och Stevens E.B. (1991) Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. Perception and Psychophysics, 50, 524-536.

Green K.P. (1996) Studies of the McGurk effect: Implications for theories of speech perception. ICSLP’96, Philadelphia, USA. http://www.asel.udel.edu/icslp/cdrom/vol3/1007/a1007.pdf

Hayashi T., och Sekiyama K. (1998) Native-foreign language effect in the McGurk effect: A test with Chinese and Japanese. AVSP’98, Terrigal, Australia. http://www.isca-speech.org/archive/avsp98/

Hietanen J.K., Manninen P., Sams M., och Surakka V. (2001) Does audiovisual speech perception use information about facial configuration? European Journal of Cognitive Psychology, 13 (3), 395-407

Johnson F.M., Hicks L.H., Goldberg T. och Myslobodsky (1988) Sex differences in lipreading. Bulletin of the Psychonomic Society, 26 (2), 106-108.

Jordan T.R. och Sergeant P. (2000) Effects of distance on visual and audiovisual speech recognition. Language and Speech, 43 (1), 107-124.

Mártony J. (1974) On speechreading of Swedish consonants and vowels. STL-QPSR, 2-3/1978, 11-32.

Massaro D.W. (1987) Speech perception by ear and by eye. Hillsdale, NJ: Erlbaum.

Massaro D.W. (1999) Speechreading: illusion or window into pattern recognition. Trends in Cognitive Sciences, 3, 8(29), 310-317.

McGurk H. och MacDonald J. (1976) Hearing lips and seeing voices. Nature, 264, 746-748.

Nilsson F. (2003) Två dialekter – två generationer. D-uppsats. Institutionen för lingvistik. Stockholms universitet.

Risberg A. och Agelfors E. (1978) Information extraction and information processing in speech-reading. STL-QPSR, 2-3/1978, 62-82.

Rosenblum L.D. och Saldaña H.M. (1996) An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 22, 318-331.

Sams M., Aulanko R., Hämäläinen M., Hari R., Lounasmaa O.V., Lu S.-T. och Simola J. (1991) Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. Neuroscience Letters, 127, 141-145.

Sekiyama K. och Tohkura Y. (1993) Inter-language differences in the influence of visual cues in speech perception. Journal of Phonetics 21, 427-444.

Strange W. (1995) Cross-language studies of speech perception: A historical review. Speech perception and Linguistic experience: Issues in cross-language research. Timonium, MD: York Press, Inc, 3-45.

Sumby W. H. och Pollack I. (1954) Visual contribution to speech intelligibility in noise. Journal of Acoustical Society of America, 26, 212-215.

Summerfield A.Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. Hillsdale, NJ: Erlbaum

Summerfield A.Q., och McGrath M. (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels. Quarterly Journal of Experimental Psychology, 36A, 51-74.

Traunmüller H. (1979) Lippenrundung bei schwedischen Vokalen. Phonetica, 36, 44-56.

Traunmüller H. (1994) Conventional, biological and environmental factors in speech communication: A modulation theory. Phonetica, 51, 170-183