UNIVERSIDAD AUTÓNOMA DEL
ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN CIENCIAS, CINC

Integración multimodal como predicciones. El caso del efecto McGurk

T E S I S

QUE PARA OBTENER EL GRADO DE

Maestría en Ciencias

PRESENTA

Marco Antonio Flores Coronado

DIRECTOR DE TESIS
Dr.Bruno Lara Guzmán

CUERNAVACA, MORELOS

Índice general

CAPÍTULO

Índice de figuras

Índice de cuadros

RESUMEN

Integración multimodal como predicciones. El caso del efecto McGurk

Marco Antonio Flores Coronado

Constantemente nos encontramos ponderando multiples fuentes de información sensorial para darle sentido al mundo. Esto es cierto también durante la comprensión del lenguaje; por ello, el efecto McGurk ha sido interpretado como un ejemplo de integración multimodal en el lenguaje.

El efecto McGurk ha sido ampliamente utilizado para estudiar la integración de la información audiovisual durante la percepción del lenguaje. El efecto ocurre cuando se parean estímulos audiovisuales silábicos incongruentes (audición: *ba* + visión: *ga* = percepto mixto: *da*)

En la presente estudiamos el efecto McGurk por medio de una red neuronal artificial (*SOIMA*) con una architectura jerárquica que aprende por procesos de autoorganización y de aprendizaje de coocurrencias. La arquitectura nos permite estudiar el efecto McGurk mediante la comparación de la integración multimodal de sus entradas con respecto de predicciones basadas en las coocurrencias audiovisuales aprendidas.

Durante la prueba de la arquitectura, comparamos la similitud entre los patrones de integración multimodal de estímulos audiovisuales congruentes e incongruentes. Nuestros análisis sugieren que los estímulos incongruentes son procesados de manera diferente a los congruentes; sin embargo, el percepto se resuelve en favor de la entrada sensorial más confiable.

Nuestros resultados sugieren que el efecto McGurk no es ocasionado por la mezcla de información audiovisual incongruente, sino por la resolución de errores de predicción multimodal. Descubrimos también que el percepto *da* no emerge cuando solo compiten *ba*, *ga* y *da*; explicamos su emergencia en humanos como consecuencia de la competencia de varias sílabas.

Capítulo 1

Introducción

Los seres vivos nos encontramos inscritos en un ambiente rico en información unimodal (vista, sonido, propiocepción, entre otros) que coocurre temporal o espacialmente (Stein and Rowland, 2019). En consecuenica, los individuos emplean las entradas sensoriales que reciben y reducen la ambigüedad perceptual del mundo a través de procesos de integración multimodal (Wallace and Stein, 1997). Estos procesos se desarrollan posnatalmente y tendrán un efecto en la eficacia con la que el individuo se percibe a sí mismo en relación con su ambiente (Wallace and Stein, 2007). El mal desarrollo de la intregración multimodal, ocasiona problemas del desarrollo en los individuos (Stein and Rowland, 2011).

El lenguaje es una facultad humana que permite la comunicación (Hauser, 2002; Hochett and Hockett, 1960) y que se asocia a habilidades cognitivas de dominio general como el aprendizaje de coocurrencias unimodales (Frost et al., 2015). El lenguaje, por otro lado, es un ambiente comunicativo humano que depende de diferentes entradas unimodales (e.g., visión, sonido, emociones, entre otras) para posibilitar la comunicación (Drijvers et al., 2019; Campbell et al., 2001); por lo tanto, la coherencia lingüística depende de procesos de integración multimodal que faciliten la percepcion lingúística y la emergencia semántica(Morse and Cangelosi, 2017).

Por ejemplo, se ha comprobado que la eficacia comunicativa del lenguaje vía auditiva se ve acrecentada cuando es acompañada por información visual (Deonarine et al., 2012); asimismo, se ha observado que los hablantes tienen la capacidad de regular su producción en relación a la retroalimentación sensorial que reciben (Sato and Shiller, 2018). Lo anterior fortalece la postura de que la comprensión lingüística reside en procesos predictivos de integración multimodal, mismos que han sido simulados a través de modelos internos (Olasagasti et al., 2015; Ursino et al., 2014; Houde and Nagarajan, 2011; Tourville and Guenther, 2011).

El aprendizaje de las asociaciones multimodales del lenguaje se determina mediante las coocurrencias de la activación de procesos unimodales (senso-

riales). La activación multimodal integra la información de las activaciones unimodales. Estas activaciones unimodales se transmiten a través de las conexiones que se reforzaron entre la información unimodal y multimodal; asimismo producen predicciones multimodales. Cuando la predicción es incongruente con la activación multimodal, se genera un error entre ambas que deberá de ser disminuido.

Esta perspectiva del aprendizaje lingüístico por coocurrencias sensoriales y procesos multimodales ha sido empleada con eficacia para simular en agentes artificiales desde la adquisición de los sonidos de una lengua, hasta el aprendizaje de palabras mediante procesos cognitivos como el mapeo rápido y la exclusividad mutua. a la vez, se observa que el desarrollo de dichas habilidades lingüísticas en agentes artificiales sufren un desarrollo en U similar al que se observa en niños e infantes (Morse and Cangelosi, 2017; Belpaeme and Morse, 2012; Morse et al., 2011; Twomey et al., 2016).

El efecto McGurk, por otro lado, es una ilusión perceptual de integración de entradas incongruentes (Mcgurk and Macdonald, 1976). El fenómeno sucede cuando ante pares audiovisuales incongruentes ($visual[ga] + auditivo/ba/$) se percibe un percepto incongruente con las entradas (tradicionalmente reportado como $[da]$). La percepción incongruente es considerada un percepto mixto y se explica como una falla en la predicción e integración multimodal de la información visual y auditiva; empero, el fenómeno no siempre ocurre pese a la existencia de información incongruente.

En la presente, empleamos una modificación del SOIMA (*Self Organized Internal Model Arquitecture*)(Escobar-Juárez et al., 2016; Lara et al., 2018) para modelar el proceso de integración audivisual durante el lenguaje. La arquitectura se entrena con diadas congruentes y es puesta a prueba con diadas audiovisuales incongruentes a fin de simular las condiciones que originan el efecto McGurk. Estudiamos además como emergencia del percepto mixto $[da]$ no se explica mediante la mezcla de información incongruente, si no mediante la reducción del error de predicción multimodal. Argumentamos que la integración multimodal del lenguaje aumenta la eficacia del discernimiento modal en condiciones naturales. También sostenemos que la falta de consistencia en percibir los estímulos McGurk como $[da]$ se debe a que dicha representación multimodal no es el resultado de la mezcla de la información modal $[ba]$ y $/ga/$, sino una de las posibilidades de interpretación a causa de la disminución del error. La manera en cómo se percibe la disminución del error depende de la experiencia lingüistica en términos de la resolución del error entre las activaciones multimodales y las predicciones multimodales.

1.1. Hipótesis

1. La aparición de perceptos mixtos ante estímulos McGurk se origina a causa de un intento de disminuir el error suscitado entre la información multimodal predicha y la real.

2. La emergencia del percepto mixto en estímulos McGurk depende de la experiencia multimodal individual.

3. Los perceptos mixtos no se originan por la mezcla de información incongruente.

1.2. Objetivo

Proponer una implementación computacional de la integración auditiva y visual que explique la emergencia del efecto McGurk como consecuencia del procesamiento multimodal del lenguaje.

1.3. Objetivos Específicos

1. Desarrollar un modelo computacional del procesamiento multimodal basado en la auto-organización y el aprendizaje estadístico de co-ocurrencias sensoriales (unimodales).

2. Proponer que, en estímulos McGurk, el percepto mixto [*da*] no es el resultado de la mezcla de los perceptos congruentes [*ba*] y [*ga*], sino una solución posible que depende de la experiencia multimodal del individuo.

Capítulo 2

Efecto McGurk

El efecto McGurk (Mcgurk and Macdonald, 1976) es un fenómeno de integración audiovisual durante la comprensión lingüística (Keough et al., 2019; Holler and Levinson, 2019; Christiansen and Chater, 2015). El efecto sucede de la siguiente manera: Se presenta un video silente con un hablante articulando $[ga]$ (ga_v), mientras que auditivamente es presentada la sílaba $/ba/$ (ba_s); la diada de información incongruente suele ser interpretada como el *percepto mixto*$[da]$. El efecto es generalizable a otras diadas incongruentes; por ejemplo, cuando se prensenta $[ka](ka_v)$ y $/pa/(pa_s)$; se interpreta como $[ta]$.

La aparición del percepto mixto es considerada como un fenómeno suficientemente robusto debido a que se ha confirmado su aparición en diferentes lenguas, culturas y durante diferentes etapas del desarrollo humano; incluso ha sido estudiado para intentar caracterizar poblaciones con problemas del desarrollo como el Trastorno del Espectro Autista, entre otros (Magnotti et al., 2015; Hirst et al., 2018; Zhang et al., 2019).

Algunos estudios han correlacionado exitosamente la aparición del percepto mixto con la activación del (STS_l) (*left Superior Temporal Sulcus*) (Beauchamp, 2016; Nath and Beauchamp, 2012; Calvert et al., 2001) lo que ha llevado a concluir que dentro de dicha área sucede la integración de la información multimodal referente al lenguaje. Paralelamente, se ha observado que la presentación de estímulos McGurk genera ondas *alpha* parecidas a las elicitadas durante una tarea tipo Stroop (Van Engen et al., 2019), lo que sugiere que la integración del percepto mixto es la manera del cerebro de resolver un problema de incongruencia multimodal; a mayor error, mayor la activación del STS_l.

Existe una diferencia en la velocidad de transmisión entre la información visual y la información auditiva que causa un retraso en la recepción sensorial de la segunda de aproximadamente 120 ms. Aprendemos a predecir información multimodal para compensar este desfase temporal y reducir ambigüidad. Por ello, se ha propuesto que la naturaleza del efecto no reside en la integración multimodal de información incongruente, sino en la incongruencia entre

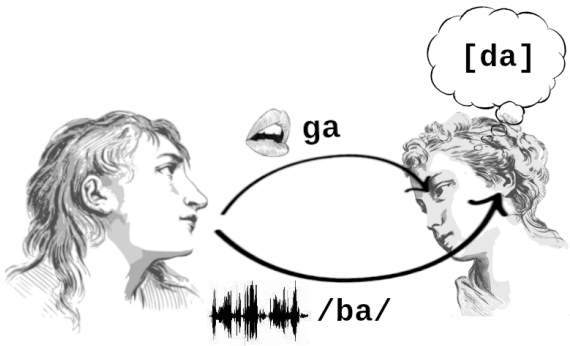

Figura 2.1: El efecto McGurk ocurre cuando se presentan simultáneamente los estímulos [*ga*] visual y [*ba*] auditivo; generalmente ocasiona el percepto mixto [*da*].

la información multimodal predicha en relación con la información existente (Olasagasti et al., 2015; Tian and Poeppel, 2012).

Debido a que la información percibida es incongruente temporalmente, se generán predicciones multimodales con base en la información disponible, aumentando así la eficiencia para percibir la información modal como multimodal. En estímulos McGurk, el error entre la comparación de la información predicha (predicción de activación multimodal) y la recibida (activación multimodal) conlleva una disminución del error que da como resultado un percepto mixto (la percepción de un estímulo que no corresponde con ninguna de las entradas).

Se ha confirmado empíricamente la emergencia del precepto mixto [*da*] cuando se presenta un adelanto de aproximadamente 200 ms entre la aparición de la información visual y la auditiva (van Wassenhove et al., 2007; Miller, 2005). Por el contrario, cuando la información auditiva antecede a la visual, la ventana de integración tolerada por los participantes suele ser de aproximadamente 100 ms (van Wassenhove et al., 2007; Miller, 2005). Es decir, existe una mayor tolerancia al desfase sensorial visual frente al auditivo probablemente debido a la diferencia natural de transmisión entre ambas entradas.

También se ha observado en estímulos McGurk que existe una correla-

ción moderada entre la habilidad de un participante para leer labios –predecir sonidos únicamente a partir de la presentación de videos insonorizados– y la probabilidad de aparición del efecto McGurk (Strand et al., 2014; Brown et al., 2018).

Todo lo anterior sirve como evidencia de la relevancia de la habilidad de decodificiación de la información modal dentro de un periodo temporal congruente para lograr la integración multimodal lingüística. Esto valida la existencia de una predisposición por integrar información temporalmente incongruente; asimismo, proporciona evidencia indirecta de que los perceptos mixtos no son ocasionados por una integración temporalmente pareada, sino por una integración multimodal entre la información audiovisual existente y la predicha dentro de un periodo temporal.

Pese a la robustez del efecto McGurk, la incidencia de este posee una gran variabilidad de aparición entre individuos. Existen sujetos que nunca reportan percibir el percepto mixto, mientras que otros siempre dicen percibirlo(Basu Mallick et al., 2015; Magnotti et al., 2018a). Esta variabilidad no se explica por condiciones intersujetos de habilidades cognitivas como la memoria de trabajo, ni por la demanda atencional propia de la tarea (Magnotti et al., 2018a; Brown and Strand, 2019); sin embargo, sí se correlaciona con la cantidad de atención que prestan los participantes a la boca (atención visual). Esta conducta robustece la cantidad de información multimodal lingüística disponible (Hisanaga et al., 2016).

La habilidad de relacionar movimientos bucales con sonidos lingüísticos es fortalecida cuando en actos de habla existe una tendencia atencional a realizar fijaciones de mirada por sobre de la zona bucal(Zhu and Beauchamp, 2017). Dicho cambio conlleva un incremento en la inteligibilidad de la información lingüística, pues disminuye la ambigüedad de la información únicamente auditiva o visual.

Se ha confirmado que parte de la variabilidad en la incidencia de aparición del efecto McGurk se debe a inconsistencias metodológicas, entre las que destacan: 1) la construcción de los estímulos (cantidad sílabica, hablantes, presentación de videos faciales completos o centrados en la boca, retraso entre modalidades, duración, entre otros); 2) la poca estabilidad entre las muestras utilizadas (tamaño de las muestras, diferencias culturales y la alta variabilidad de lenguas maternas de los participantes); 3) las discrepancias en la forma en que los participantes reportan sus respuestas (de respuesta forzada o abierta); 4) la discordancia entre estudios para determinar criterios de clasificación de las respuestas (Van Engen et al., 2019; Magnotti et al., 2018b; Basu Mallick

et al., 2015; Magnotti et al., 2018a).

Existen varias críticas sobre la interpretación de la integración multimodal lingüística y del efecto McGurk en sí (Van Engen et al., 2019; Basu Mallick et al., 2015; Keough et al., 2019,0). Por ejemplo, se ha descubierto mediante *fMRI* (*Functional Magnetic Resonance Imaging*) que la sensibilidad al percepto mixto depende del estado de actividad encefálico previo a la aparición de los estímulos McGurk (Basu Mallick et al., 2015). Por otro lado, se ha argumentado que el lenguaje no se codifica únicamente a través de las modalidades auditiva y visual, sino que muchas otras particicipan en la integración lingüística (como la somatosensorial) (Van Engen et al., 2019; Keil et al., 2012). Asimismo, se ha criticado que la mayoría de los estudios califique como percepto mixto únicamente a las respuestas de tipo [*da*] o [*ta*], mientras que el resto de perceptos son catalogados como "errores" de interpretación de los sujetos (Van Engen et al., 2017). Esta calificación implica a priori que [*da*] no es un error de interpretación pues es la "mezcla" de información entre [*ba*] y [*ga*].

En contraposición a la anterior, existe evidencia de que presentarr reiteradamente estímulos McGurk conlleva un reajuste de los límites acústicos y artículatorios de los estímulos visuales, auditivos y mixtos orillándolos a identificar sílabas con mayor flexibilidad (Gentilucci and Cattaneo, 2005; Nahorna et al., 2015; Lüttke et al., 2016). La interpretación audiovisual es flexible y se encuentra en constante ajuste. La sensibilidad a reportar el percepto mixto depende del tipo de estímulo anterior. Este reajuste se observa también en que los sujetos producen una sílaba /*ba*/ menos parecida a su linea base y más cercana acústicamente a /*da*/ después de presenciar estímulos McGurk.

Por otro lado, también se ha observado con electroencefalografía una reducción en la elicitación del MMN (*Mismatch Negativity*) tras haber presenciado estímulos McGurk (Lüttke et al., 2016); el MMN es un potencial sensible a disparidades auditivas. Es decir, existe evidencia electroencefálica de que los estímulos McGurk progesivamente se perciben más similares a alguna de las entradas unimodales. Desde nuestra perspectiva, dicho fenómeno sucede porque tras la aparición de un estímulo McGurk se actualizan las predicciones multimodales mediante la nueva coocurrencia modal y se resuelve progresivamente mejor el error en favor de alguna de las entradas unimodales.

2.1. Implementaciones Computacionales

La mayoria de los trabajos enfocados en simular y estudiar la emergencia del efecto McGurk son modelos probabilístios que, aunque resultan eficaces en la elicitación del efecto McGurk como [*da*], asumen a priori que la integración multimodal de [*ba*] y [*ga*] es el igual a la mezcla entre ambas entradas (Olasagasti et al., 2015; Ursino et al., 2014; Houde and Nagarajan, 2011; Tourville and Guenther, 2011). Bajo dicha premisa, se codifican los estímulos o la interacción entre la información modal y por ello el único percepto mixto posible por elicitar en mayor o menor grado es [*da*]. Esta presuposición teórica ocaciona aproximaciones operacionales deficientes para explicar la variabilidad de aparición del efecto McGurk. Estos acercamientos, además, proponen procesamientos específicos para la integración audiovisual lo que es incongruente con la perspectiva de que el lenguaje es una habilidad de dominio general (Frost et al., 2015).

Pese a lo anterior, existen dos acercamientos computacionales que al contrario de los anteriores, se basan en la coocurrencia y la autorganización perceptual (Omata and Mogi, 2008; Gustafsson et al., 2014). Estos modelos, además, no presuponen operacionalmente a [*da*] como el único percepto posible.

Omata and Mogi (2008) emplean un SOM (*Self-Organized Map*)que entrenan con vectores que contienen tanto la información visual y auditiva de varias sílabas de videos reales. Los autores confirman que la información audiovisual aumenta la inteligibilidad modal y descubren que los estímulos McGurk elicitan diversas respuestas ([*da*],[*bga*],[*gba*]); sin embargo, presuponen que la información unimodal se integra directamente como multimodal en un solo mapa. Es decir, no tratan a ambas entradas como modalidades diferentes y excluyen así el aprendizaje de coocurrencias unimodales que precede a la integración multimodal.

Gustafsson et al. (2014) emplean SOMs modales para procesar la información auditiva y visual de manera independiente y luego la asocian en un SOM multimodal logrando una arquitectura biologicamente plausibe que se basa en la autoorganización y no depende de reglas a priori para reducir el error ni para mezclar información unimodal. En su modelo, la posición y el nivel de activación de las neuronas ganadoras en los SOMs unimodales sirven de input para el SOM multimodal; esto genera un error y la información fluye en ciclo únicamente a la modalidad auditiva para reforzar la información auditiva en favor de la multimodal.

Aunque Gustafsson et al. (2014) simulan con éxito la emergencia del per-

cepto mixto, no consideran que la coactivación entre los elementos unimodales y la percepción multimodal genera activaciones entre sus elementos; por demás, asume que para reducir el error, es necesario reforzarar únicamente la activación el SOM auditivo con el output del multimodal lo que causa que se favorezca la información auditiva por sobre la visual de manera artificial. El modelo asume que la fidelidad de la percepción auditiva es la que orilla el percepto mixto en favor de uno u otro elemento y no la visual; además, implica que la reducción del error es sensorial y no perceptual.

Nuestro acercamiento posee varias ventajas frente a los modelos anteriores:

a) Confía únicamente en la autoorganización y la coocurrencia para desarrollar aprendizaje.

b) Utiliza y procesa estímulos reales.

c) Procesa el error dentro de la red multimodal.

d) Considera las represetaciones lingüísticas como multimodales.

Capítulo 3

Método

3.1. Estímulos

Se utilizaron 24 videos diferentes capturados en condiciones controladas y especialmente desarrollados para estudiar experimentalmente el efecto McGurk (Basu Mallick et al., 2015) (8 videos por sílaba de diferentes hablantes). Los videos se encuentran a color, en formato .mp4 y son de duración variable —8 hablantes diferentes (4 mujeres, 4 hombres) en vídeos congruentes de las sílabas /*ba*/, /*da*/ y /*ga*/ —. Los estímulos se encuentran disponibles en la base de datos de acceso libre especializada en el Efecto McGurk del *Beauchamp Lab* (Basu Mallick et al., 2015).

A continuación, se detallarán los procesos empleados para la extracción y procesamiento de la información auditiva y visual que conforman los videos. Los procesos de extracción de la información auditiva y visual consideran la fortaleza de la visión en el procesamiento espacial, así como la de la audición para procesar información temporal (Frost et al., 2015; Conway and Christiansen, 2005).

El proceso para caracterizar la información visual se basa en la identificación de regularidades en la dirección y el movimiento de los labios (Figura 3.1), mientras que la caracterización auditiva se centra en el comportamiento de la información acústica a través de tiempo (ver Figura 3.2).

3.1.1. Estímulos visuales

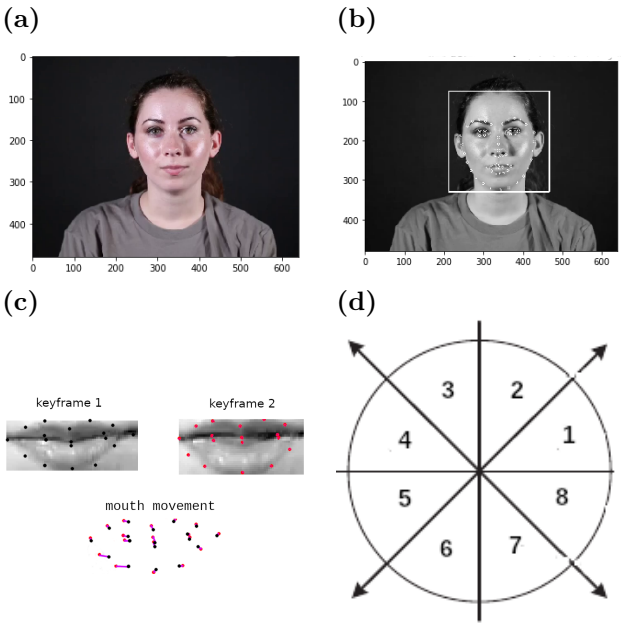

Figura 3.1: Proceso de caracterización de la información visual por tipo de sílaba. **(a)** video original en RGB; **(b)** se reducen los canales por cuadro y se identifica el rostro y los 20 puntos bucales que servirán para determinar la *ROI*; **(c)** se identifican cuadros clave; **(d)** se calcula la dirección y la magnitud del desplazamiento de cada punto bucal entre cuadros clave.

El procesamiento de la información visual considera información aspectual y geométrica para la elaboración de los descriptores. Los descriptores son una caracterización de la dirección y el movimiento de 20 puntos bucales (14 para para la parte externa de los labios y 6 para la parte interna). Los descriptores finales son vectores con 20 dimensiones y se extrajeron de la siguiente manera:

1. Se redujeron los canales de color de los videos de rgb a escala de grises
2. Se identificó el rostro dentro del video (Viola and Jones, 2001; Lienhart and Maydt, 2002).

3. Se identificaron 20 puntos bucales en cada rostro (Kazemi and Sullivan, 2014). La zona que contiene los puntos se considera nuestra *ROI* (*Region of Interest*).
4. Se extrajeron cuadros clave a partir del cambio aspectual de la *ROI* a través del tiempo; si la correlación entre el cuadro$_t$ y el cuadro$_{t+i}$ es >= 0,5 & <= 0,65, entonces se considera cuadro clave al $_{t+i}$ –a partir de entonces se considera al nuevo cuadro clave como el cuadro$_t$.
5. Se construye un *OH-ROF*(*Oriented Histograms of Regional Optic Flow*): Se generan histogramas de 8 bins por punto bucal, cada bin representa cambios en la orientación en rangos de 45° (Liu et al., 2016). Como resultado, entre cada par de cuadros clave se construye un histograma por cada punto bucal de la *ROI* (p.e., si un punto bucal X se desplaza 3 unidades con una orientación igual a 43°, entre el cuadro$_t$ y el cuadro$_{t+1}$ el histograma resultante de movimiento-dirección es $(0, 3, 0, 0, 0, 0, 0, 0)$.
6. Se acumula la información de los histogramas generados por cada punto bucal y se concatenan los 20 histogramas en un descriptor de 160 unidades normalizado entre 0 y 1.

El procesamiento anterior genera un vector por sujeto con una longitud igual a 160 unidades.

A razón de aumentar la cantidad de estímulos disponibles se realizó el siguiente procedimiento:

1. Se asumió que la caracterización del movimiento bucal por clase de sílaba pertenece a una distribución multivariable donde cada dimensión se obtiene de la distribución normal de cada histograma de flujo regional óptico
2. Se emplearon 4 Cadenas de Montecarlo de 100,000 elementos cada una (Hoffman and Gelman, 2014) para muestrear la multivariable; las primeras 35,000 muestras de cada cadena fueron desechadas para ajustar el algoritmo. Lo anterior resultó en un conjunto final de 400,000 elementos pertenecientes a la distribución conjunta multivariada de la articulación de una sílaba.
3. Se seleccionaron los 2,000 estímulos con mayor densidad de función del conjunto total por ser los más probables.
4. Se reordenó aleatoriamente la submuestra y se dividió en 11 subconjuntos de 181 muestras por clase de sílaba (10 subconjuntos servirán

como entrenamiento de diferentes SOIMAs, el restante es utilizado como subconjunto de prueba).

3.1.2. Estímulos auditivos

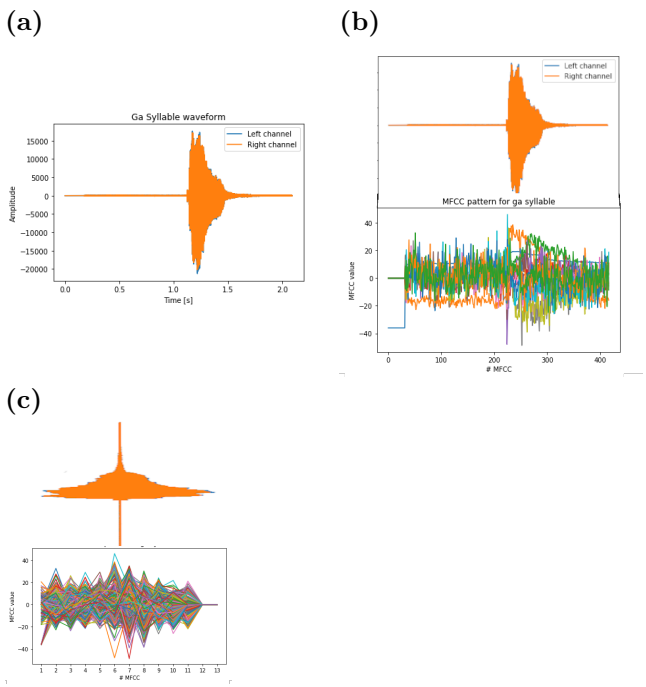

Figura 3.2: Proceso de extracción de información auditiva por tipo de sílaba. **(a)** Se aisló la información auditiva del video fuente; **(b)** se extrajeron 13 MFCC por ventana de análisis; la cantidad de ventanas finales dependía de la duración del audio. **(c)** Se calculó la distribución de cada MFCC a través de las n ventanas de análisis

.

El preprocesamiento de la información auditiva se basa en el uso de MFCCs (*Mel-Frequency Cepstral Coefficients*) para caracterizar el habla (Zhen et al., 2000; Gold et al., 2011). Los descriptores auditivos finales describen los cambios que sufren los elementos dentro de cada MFCC a lo

largo del tiempo. Los descriptores finales son vectores de 13 dimensiones y se extrajeron de la siguiente manera:

1. Se extrajo la información sonora del video como formato WAV. Los audios resultantes son de longitud variable.
2. Se determinaron ventanas de análisis de 25 ms de duración con 1 ms de separación entre ellas. A cada una de las ventanas se le aplicaron 512 Transformaciones Rápidas de Fourier (FFT) y se extrajeron 13 MFCC por ventana. Los vectores resultantes poseen una longitud variable dependiendo de la longitud del audio.
3. Se asumió que cada MFCC pertenece a una distribución independiente, a fin de caracterizar las regularidades temporales y obtener vectores de longitud invariante grupales por sílaba,; el resultado final es una caracterización grupal por sílaba de 13 gaussianas.

A razón de aumentar la cantidad de estímulos disponibles se realizó el siguiente procedimiento:

1. Se asumió que la caracterización auditiva silábica es una distribución multivariable.
2. Se emplearon 4 Cadenas de Montecarlo de 100,000 elementos cada una para muestrear la multivariable (Hoffman and Gelman, 2014); las primeras 35,000 muestras de cada cadena fueron desechadas para ajustar el algoritmo. Esto resultó en un conjunto final de 400,000 elementos pertenecientes a la distribución conjunta multivariada del comportamiento sonoro de una sílaba.
3. Se seleccionaron los 2,000 estímulos con mayor densidad de función por ser los más probables.
4. Se reordenó aleatoriamente a la muestra y se dividió en 10 subconjuntos de 181 muestras por clase de sílaba.

3.1.3. Arquitectura

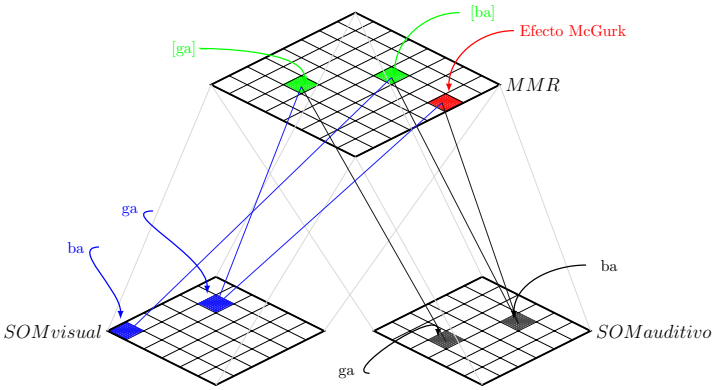

Figura 3.3: Diagrama de la arquitectura propuesta. El flujo de la activación modal se presenta en color verde, el flujo de la predicción de activación multimodal se presenta en color azul. En color rojo, se presenta la activación corregida que da como resultado el percepto usualmente asociado a [*da*]

Para modelar la integración audiovisual del lenguaje se adaptó la arquitectura SOIMA (*Self-Organized Internal Model Architecture*)(Escobar-Juárez et al., 2016) (Figura 3.3). El SOIMA es una estructura biológicamente plausible basada en SOMs que permite simular la integración de información modal en representaciones multimodales.

Los SOMs son redes neuronales artificiales que, una vez entrenados, realizan una organización topológica de baja dimensionalidad de los estímulos con que fue entrenada –2 o 3 dimensiones– (Kohonen, 1982). La organización del SOM explota las similitudes geométricas que existen entre los elementos de entrenamiento y posee la ventaja de que no requiere de conocimiento previo sobre los estímulos que procesa (aprendizaje no supervizado).

Tras el entrenamiento, cada neurona del SOM posee un valor prototípico con base en el conjunto de entrenamiento dada la distribución topológica del mapa. Para la presente investigación, entrenamos 10 SOIMAs con diferentes subconjuntos de entrenamiento para reproducir parte de la variabilidad en la resolución de problemas de integración de informaicón

incongruente.

Las modificaciones del SOIMA se realizaron considerando trabajos anteriores en los que se modelaron con éxito procesos de adquisición del lenguaje en el robot i-Cub con la arquitectura ERA (Morse et al., 2011; Twomey et al., 2016; Belpaeme and Morse, 2012; Morse and Cangelosi, 2017).

Coincidimos con los autores de dichas implementaciones en que existe una activación multimodal marcada por la información proveniente de las zonas modales, así como una propagación de la activación que fluye desde los SOMs modales y se origina por el aprendizaje estadístico de las coocurrencias entre los elementos unimodales y el multimodal; sin embargo, consideramos que su acercamiento no es biológicamente plausible, pues para hallar la convergencia entre los dos tipos de activaciones, la información debe de saltar de un SOM a otro varias veces. Lo anterior presupone que para percibir lenguaje deben de realizarse múltiples predicciones bottom-up y top-down.

Para resolver la integración multisensorial de la información visual y de la información auditiva durante la producción de sílabas, empleamos un SOM para procesar la información visual que simula el procesamiento modal de las cortezas visuales (SOM_V) y un SOM para procesar la información auditiva de las sílabas que simula el proceamiento modal de las cortezas auditivas (SOM_A). También utilizamod un SOM multimodal (MMR) que símula el procesamiento funcional de asociación multimodal que ocurre en el STS_l.

Entrenamiento SOMs

La implementación del SOIMA requiere de diversos procesos off-line para caracterizar y extraer la información auditiva y visual del contenido multimedia origen (véase *Estímulos*). En la ecuación 3.1 presentamos la función de activación que utilizamos para los mapas. En cada SOM se utiliza como función de activación al valor inverso de la distancia euclideana normalizada entre la entrada y el mapa .

$$Act_j = \sqrt{\sum (v_i - w_j)^2} \tag{3.1}$$

Donde Act_j es la activación de la neurona j del SOM tras el procesamiento del vector de entrada v_i; w_j es el vector de pesos que posee la neurona j. Tras calcular la distancia euclideana de cada neurona con respecto de la entrada, aplicamos una normalización MinMax sobre el conjunto de activaciones; esto da como resultado que la menor distancia euclideana obtenga un valor igual a 0, mientras que la mayor distancia euclideana obtenga un valor igual a 1. Finalmente, utilizamos el valor inverso del valor de activación normalizado. Este proceso da como resultado que la neurona con una distancia normalizada igual a 0 posea una activación final igual a 1; mientras que otra neurona con una distancia normalizada de 1 posea una activación final igual a 0. La neurona ganadora (BMU) es la neurona j con una activación final igual a 1.

La entrada del SOM_V es un vector de 20 dimensiones; la entrada del SOM_A es un vector de 13 dimensiones; y la entrada del MMR es un vector de cuatro dimensiones con las coordenadas de las neuronas ganadoras de SOM_V y SOM_A normalizado entre 0 y 1 (e.g., si poseemos 2 mapas modales cuadrados con 6 neuronas por lado, y si la neurona ganadora de SOM_V se encuentra en $(0, 0)$, mientras que la neurona ganadora de SOM_A se encuentra en $(5, 5)$, la entrada que le corresponde al MMR es $[0, 0, 1, 1]$).

En la Ecuación 3.2 describimos el proceso mediante el cual actualizamos los pesos de cada neurona del SOM una vez que la respectiva BMU es identificada.

$$\Delta w_{ij} = (\alpha\lambda)h_j(v_i - w_j) \tag{3.2}$$

Donde Δ w_{ij} es el incremento en el vector de pesos de cada neurona j del mapa, con respecto de la entrada i; w_j es el vector de pesos original de la neurona j del SOM, v_i es la entrada; h_j es la función de vecindad sobre la neurona j. Por otro lado, α es la tasa de aprendizaje (0.3) y λ es la proporción de ciclos de entrenamiento restantes, esto se decidió para fomentar que la tasa de aprendizaje descienda linealmente de 0.3 a 0.

En la Ecuación 3.3 describimos la función del vecindario de aprendizaje que empleamos. Se puede observar que el tamaño del vecindario desciende linealmente a lo largo del entrenamiento. Para lograr esto, la función calcula un gaussiana centrada en la BMU.

$$h_j = e^{\frac{-\beta_j}{2\sqrt{n^2}}} \tag{3.3}$$

Donde h_j es el tamaño del radio de vecindad de aprendizaje alrededor de la BMU. β_j es la distancia entre la neurona j y la BMU. Si β_j es mayor que el tamaño del vecindario de aprendizaje, $h_j = 0$. El tamaño del vecindario decrese linealmente desde v_i = al tamaño del mapa, hasta v_f = 1. La función nos garantiza que al inicio del entrenamiento el tamaño del vecindario de aprendizaje abarque todas las neuronas del SOM y que el tamaño de vecindario de aprendizaje disminuya progresivamente durante el entrenamiento hasta afectar únicamente a la BMU.

Aprendizaje multimodal

En nuestra implementación, para simular la asociación multimodal entre los SOMs modales a través de la cual fluyen las predicciones, utilizamos la regla de aprendizaje hebbiano que postula lo siguiente:

> Cuando el axón de una célula A está lo suficientemente cerca como para excitar a una célula B y repetidamente toma parte en la activación, ocurren procesos de crecimiento o cambios metabólicos en una o ambas células de manera que tanto la eficiencia de la célula A, como la capacidad de excitación de la célula B son aumentadas.
>
> (Hebb, 1962)

Denominaremos aprendizaje hebbiano a la asociación que se da entre la BMU de cada mapa modal y la BMU del MMR.

Tras identificar la ubicación de las respectivas neuronas ganadoras de cada uno de los SOMs, utilizamos la regla de Oja (Ecuación 3.4) para modelar el aprendizaje hebbiano entre cada SOM modal y el MMR (Oja, 1982). El proceso resulta en que las neuronas del SOM_V y del SOM_A estén totalmente conectadas al MMR.

$$\Delta W_{Xy} = \alpha_h y(X - yW_{Xy}) \tag{3.4}$$

Donde $\Delta\ W_{Xy}$ es el incremento de los pesos hebbianos durante la asociación de las neuronas X de un mapa y cada neurona ganadora y del MMR. α_h es la tasa de aprendizaje hebbiano (0.08); X es un vector con las activaciones de un mapa modal con respecto de su entrada; y es la activación de cada neurona y del MMR con respecto de su entrada multimodal; W_{Xy} los pesos de asociación hebbiana entre el vector X y la neurona y. Los pesos hebbianos se inicializan aleatoriamente y se autorganizan a lo largo del entrenamiento.

Predicciones multimodales

Las predicciones multimodales nos informan sobre el tipo de estímulo multimodal que el MMR espera percibir con base en la información sensorial disponible; por ejemplo, la información visual de la sílaba [*ba*] predice la información auditivia de [*ba*] y viceversa. La predicción del MMR se basa en el aprendizaje de coocurrencias. Para calcular las *predicciones multimodales* utilizamos la Ecuación 3.5. Esta ecuación nos permite unificar las predicciones visuales y auditivas, de tal forma que el estímulo audiovisual predicho sea el más congruente con la información sensorial disponible.

$$PredAct_j = (\zeta * w_{j|BMU_A}) + (\zeta * w_{j|BMU_V}) \tag{3.5}$$

Donde $PredAct_j$ es la activación predicha de la neurona j del MMR en consecuencia de las entradas modales; $w_{j|BMU_A}$ es el valor del peso hebbiano entre la neurona ganadora A del SOM auditivo y la neurona j del MMR; $w_{j|BMU_V}$ es el valor del peso hebbiano entre la neurona ganadora V del SOM visual y la neurona j del MMR; ζ es un coeficiente de relevancia de la información modal que garantiza que cada predicción tenga la misma relevancia para el computo de la predicción de activación multimodal. El valor de ζ se decidió como el resultado de dividir 1 entre la cantidad de predicciones modales que existen; como nuestro SOIMA cuenta sólo con 2 entradas modales, $\zeta = 0.5$. El máximo valor de predicción de activación multimodal que puede obtener una neurona es igual a 1. Para calcular la predicción de activación multimodal, es necesario

aplicar previamente una función de normalización MinMax sobre todos los pesos que unen a BMU_A con el MMR, así como sobre todos los pesos que unen a BMU_V con el MMR. Esto da como resultado que el mayor peso hebbiano entre la unidad ganadora de cualquier mapa modal y el MMR posea un peso normalizado igual a uno, mientras que el menor peso hebbiano entre la unidad ganadora de cualquier mapa modal y el MMR posea un peso normalizado igual a cero.

Finalmente, para procesar el error que emerge entre Act (Ecuación 3.1) y $PredAct$ (Ecuación 3.5); y para calcular la *activación corregida* del MMR utilizamos la Ecuación 3.6. Esta ecuación nos permite calcular la unidad perceptual ganadora dentro del MMR como aquella que mejor reduce el error.

$$CorrAct_j = 1 - (\eta Act_j) + (\eta PredAct_j) \tag{3.6}$$

Donde $CorrAct_j$ es la activación corregida de la neurona j del MMR con respecto del error. η es el coeficiente de mezcla (η=0.2) y representa la relevancia que posee la *activación modal* y la *predicción multimodal* para el cálculo y resolución del error predictivo; el valor que le otorgamos a η fomenta que el SOIMA preste mayor relevancia a la información sensorial que a las predicciones de activación multimodal. Act_j es la activación modal de la neurona j del MMR que se obtiene como consecuencia de la ecuación 3.1. $PredAct_j$ es la predicción de activación multimodal de la neurona j del MMR que se obtiene como consecuencia de la ecuación 3.5.

Cuando la activación multimodal y la predicción de activación multimodal son iguales, no existe una diferencia entre $CorrAct_j$ y Act_j; sin embargo, si Act_j no coincide con $PredAct_j$, entonces $CorrAct_j$ y Act_j no son iguales. Finalmente, $CorrAct_j$ nos permite identificar a la neurona ganadora percepto (BMU_p) como aquella neurona con mayor valor de activación corregida. BMU_p se utilizó para los análisis realizados en la sección 4.2, mientras que $CorrAct_j$ para los análisis realizados en la sección 4.1.

Capítulo 4

Resultados

A continuación se enlistan los resultados obtenidos con 10 modelos entrenados durante 20,000 iteraciones con una tasa de aprendizaje (α) = 0.3 para los SOMs y de 0.08 para el aprendizaje hebbiano. Todos los SOMs empleados son mapas cuadrados con 6 unidades por lado. Los estímulos de prueba consistían en las permutaciones posibles entre las 3 clases de estímulos, lo que resultó en 3 clases de estímulos congruentes y 6 incongruentes (incluyendo estímulos McGurk) Tabla 4.1.

Cada SOIMA fue entrenado con una conjunto de datos de entrenamiento diferente para aumentar la variabilidad entre los SOIMA; todos fueron probados con un paradima de validación cruzada (ejemplo de matrices de confusion de un SOIMA entrenado, ver Figura 4.1).

4.1. Resultados (activación corregida)

Con el objetivo de determinar a qué estímulo congruente se asemeja la activación de los estímulos de prueba incongruentes, se decidió medir la similitud entre los patrones de activación corregida de los estimulos audiovisuales de prueba incongruentes con respecto de los estímulos de audiovisuales de prueba congruentes con la Ecuación 4.1. Para ello, utilizamos la *MI* (*Mutual Information*) como medida de similitud entre la activación corregida de los estímulos audiovisuales incongruentes de prueba , con respecto de sus correspondientes estímulos audiovisuales congruentes de prueba (Ver Tabla 4.2).

La MI es una medida bidireccional entre 2 distribuciones que cuantifica la cantidad de información que ambas comparten –en términos de entropía– (ver Figura 4.2). La variante de MI utilizamos corrige la información compartida como consecuencia del azar y nos ofrece valores normalizados

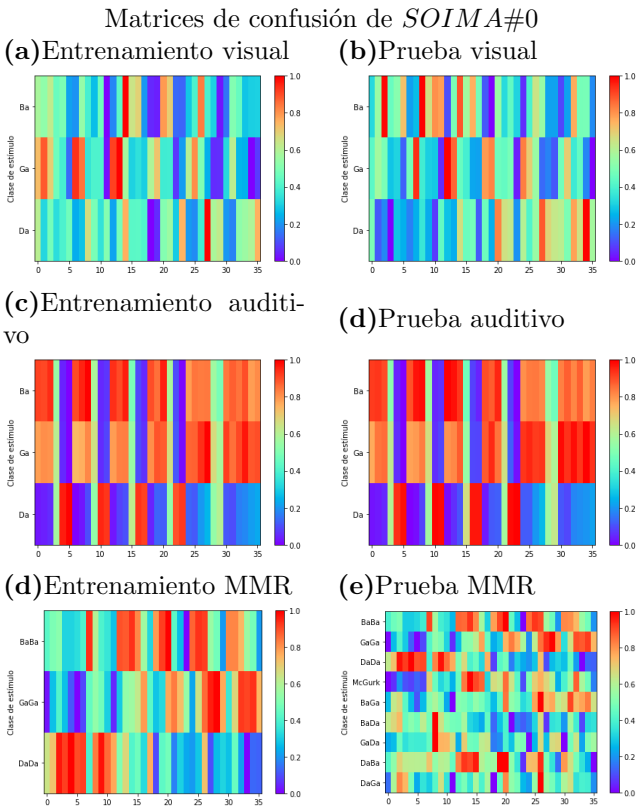

Figura 4.1: Matrices de confusión de los patrones de activación corregida por clase de estimulo. En SOM_A se observa que $[ba]$ y $[ga]$ generan activaciones indistingibles. En SOM_V se observa que las 3 clases generan activaciones distingibles. En MMR se observa que las 9 clases de estímulos multimodales (3 clases congruentes + clase McGurk + 5 clases incongruentes) generan activaciones totalmente distingibles entre sí; lo anterior aporta evidencia en contra de que los estímulos McGurk generan una activación igual a $[da]$

Cuadro 4.1: Clase y tipo de estímulos de prueba para cada uno de los 10 diferentes SOIMA

Tipo	Visual	Auditivo	Estimulo multimodal	n
congruente	Ba	Ba	**Ba**	181
congruente	Ga	Ga	**Ga**	181
congruente	Da	Da	**Da**	181
incongruente	Ga	Ba	McGurk	181
incongruente	Ba	Ga	BaGa	181
incongruente	Ba	Da	BaDa	181
incongruente	Ga	Da	GaDa	181
incongruente	Da	Ba	DaBa	181
incongruente	Da	Ga	DaGa	181

entre 0 y 1 (Vinh et al., 2010). La información mutua corregida se calculó de la siguiente manera:

$$MI_c = \frac{MI - MI_e}{Max(MI) - MI_e} \tag{4.1}$$

Donde MI_c corresponde al valor de información mutua corregida contra el azar; MI corresponde a la información mutua entre 2 distribuciones y MI_e corresponde a la información mutua esperada como consecuencia del azar. MI_c puede obtener valores entre 0 y 1, donde 0 significa que dos distribuciones no comparten información alguna y 1 significa que ambas distribuciones son idénticas (gráfica descriptiva de los diez SOIMAs en Figura 4.3)

A fin de determinar si el tipo de estímulo incongruente afecta su similitud con respeto de los estímulos congruentes, se construyó un modelo de efectos mixtos con la siguiente fórmula:

$$MI = Input * Condition + (1|SOIMA) \tag{4.2}$$

Donde MI es la variable dependiente, $Input$ y $Condition$ son los efectos fijos y $(1|SOIMA)$ es el efecto aleatorio atribuible al SOIMA. Los análisis se realizaron con los paquetes lmer4 y afex de R.

El análisis demostró que los efectos fijos $Input$ y $Condition$ generaban pendientes significativas para todos los niveles, sin embargo, ninguna interacción resultó significativa (ver Tabla 4.3).

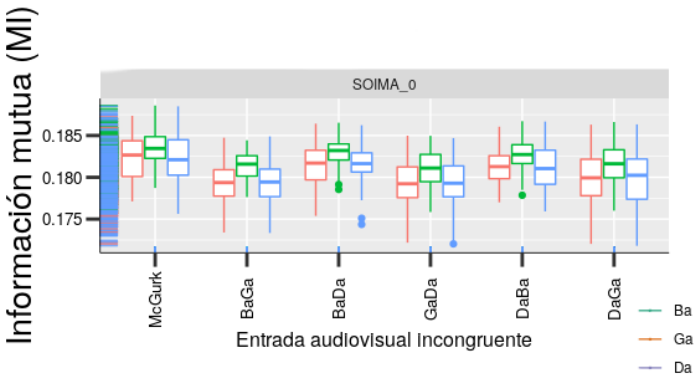

Figura 4.2: Valores de MI entre las activaciones de los estímulos congruentes e incongruentes. Los valores de MI son menores a 0.2, lo cual evidencia de que las activaciones de los estímulos congruentes e incongruentes comparten poca información entre sí. Los estímulos McGurk generan activaciones más semejantes a *ba* que a *ga* o *da*. BaGa es el estímulo incongruente inverso a los estímulos McGurk, sin embargo ambos estímulos generaron activaciones con diferentes MI.

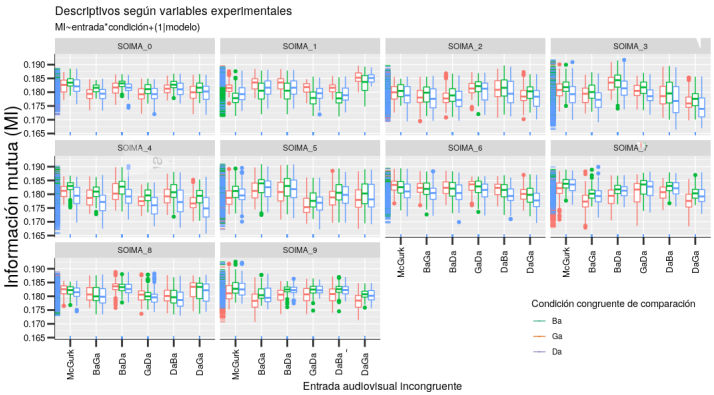

Figura 4.3: Gráfico por SOIMA de los valores de MI obtenido entre los tipos de estímulos incongruentes con respecto de los congruentes.

Cuadro 4.2: Pares de estímulos audiovisuales en los que se calculó la Información Mutua

Estimulo incongruente	Estimulo congruente	SOIMA	muestras
McGurk	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
BaGa	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
BaDa	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
GaDa	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
DaBa	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
DaGa	Ba (n = 181)	10	1,810
	Ga (n = 181)	10	1,810
	Da (n = 181)	10	1,810
Total			32,580

Cuadro 4.3: Resultados significativos para los efectos fijos del análisis de efectos mixtos

Efecto Fijo	β	Error Estándar	p
Intercepto	0.1	0.0001	< 0.001
McGurk	-0.002	0.00008	< 0.001
BaGa	-0.001	0.00008	< 0.001
BaDa	-0.001	0.00008	< 0.001
GaDa	-0.001	0.00008	< 0.001
DaBa	-0.001	0.00008	< 0.001
Ga	-0.0002	0.00008	0.002
Da	-0.0004	0.00008	< 0.001

Efectos del modelo del efectos mixtos

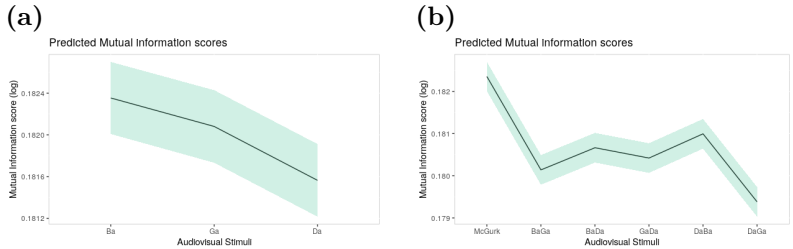

Figura 4.4: nota. Resultados significativos para los efectos fijos *Input* y *Condition*. **(a)** Evidencia que en las activaciones corregidas de los estímulos incongruentes son muy disímiles de las congruentes, sin embargo se parecen más a *Ba* y menos a *Da*; **(b)** Evidencia que de los estímulos incongruentes, los McGurk son los que más parecidos a las activaciones congruentes y permiten predecir la aparición de perceptos mixtos con el estímulo *DaBa*. La falta de interacción entre efectos fijos, nos demuestra que ningún estímulo incongruente posee una tendencia a parecerse particularmente más a alguno de los congruentes.

El análisis nos demuestra que los estímulos incongruentes se parecen en general más a *Ba* y son más disímiles de *Da*. También nos permite observar que los estímulos McGurk son el estímulo incongruente más parecido a los congruentes, sin embargo la falta de interacción sugiere que no existe una tendencia a que sus activaciones sean más parecidas a algún estímulo congruente (ver Figura 4.4).

El análisis ofrece suficiente evidencia para afirmar que a nivel de la *activación corregida* del MMR, los estímulos de tipo McGurk mantienen el mismo rango de disimilitud que cualquier otro estímulo incongruente con respecto de los estímulos congruentes; es decir, parece no existir evidencia de que los estímulos McGurk sean diferenciables con respecto de los demás estímulos incongruentes en términos de semejanza a los patrones de activación corregida de los estímulos congruentes. Finalmente, podemos concluir que la activación corregida que ocasionan los estímulos McGurk es igualmente disimil con respecto de [*da*], [*ba*] y [*ga*].

Ejemplo del proceso de etiquetado del MMR en un SOIMA

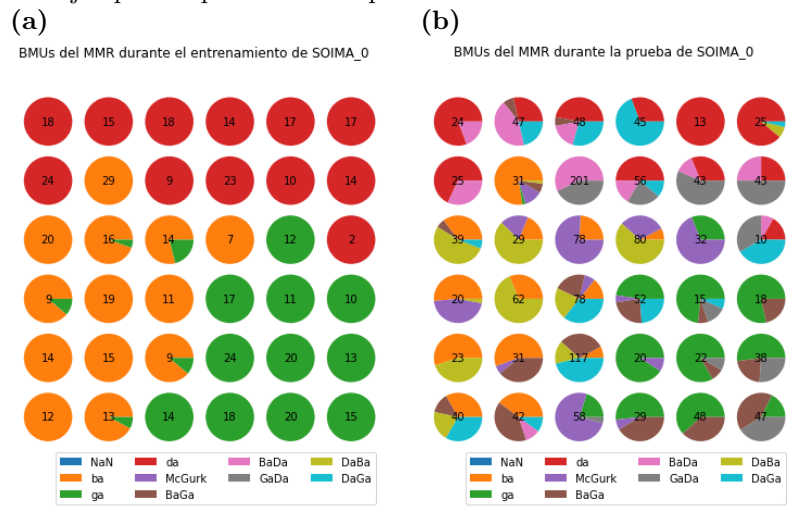

Figura 4.5: **(a)** Etiquetado topológico del MMR según la localización de la BMU por tipo de estímulo del conjunto de entrenamiento (*mapa de etiquetas*). En negro, dentro de cada unidad, la cantidad de veces que la unidad fue BMU. La etiqueta que se le asigna a cada unidad es aquella que proporcionalmente le fue más asociada (p.e., la unidad 2,2 se etiqueta como "Ba", porque fue la etiqueta que más se le asignó en el entrenamiento).**(c)** Etiquetado topológico del MMR según la localización de la BMU por tipo de estímulo del conjunto de prueba. Posteriormente, se identifican y cuentan las etiquetas que le corresponden a las BMU de acuerdo al *mapa de etiquetas* obtenido en (c).

4.2. Resultados Mejor Unidad Ganadora (BMU_p)

Para explorar la respuesta conductual de los SOIMAs frente a estímulos incongruentes, realizamos un análisis de Mejor Unidad Ganadora, donde consideramos que la respuesta perceptual de los SOIMAs era igual a la etiqueta asociada a la unidad del MMR con el mayor nivel de *activación corregida*.

Para lograr lo anterior, etiquetamos topológicamente el MMR de cada SOIMAs asociando la etiqueta del estimulo audiovisual de entrenamiento con su respectiva BMU dentro del MMR. La etiqueta que se le asignó a cada unidad es igual a la etiqueta que proporcionalmente fue más asociada a dicha unidad. Posteriormente, exploramos la distribución topológica de las BMU de todos los estímulos de prueba (ver Figura 4.5 para un ejemplo del proceso).

Dado que deseábamos conocer qué etiqueta congruente fue asociada más frecuentemente a la BMU de los estímulos tipo McGurk, realizamos una X^2 con las variables $SOIMA$ y $etiquetaBMU$. El análisis demostró que entre los 10 SOIMAs sólo existieron perceptos Ba y Ga; empero la distribución de $etiquetaBMU$ no era homogénea (X^2 =51.72, df = 9, p <0.001).

En consecuencia a lo anterior, realizamos a manera de análisis exploratorio, una prueba Kruskal-Wallis con la proporción de BMU como variable dependiente y *etiqueta* como variable independiente (X^2 =5.49, df = 1, p =0.01); los resultados fueron congruentes con el análisis previo. Una análisis post-hoc con la prueba de Mann-Whitney U fue realizada para determinar si existían diferencias en la proporción de etiquetas obtenidas frente a estímulos McGurk (Ba, Ga) (ver Figura 4.6).

Los resultados indican que existen diferencias entre Ba y Ga ($z = 19, p =$ 0.021). Estos resultados nos sugieren que en los SOIMAs los estímulos McGurk fueron asociados Ba y Ga, es decir se resolvió el error multimodal en favor de las señales modales, es decir, las respuestas siempre se asociaron a alguno de los componentes unimodales de los estímulos McGurk ("Ba" auditivo + "Ga" visual). Esto apoya nuestra propuesta de que durante la integración multimodal la mezcla de la información unimodal incongruente no basta para explicar la aparición de perceptos

mixtos.

Concluimos que la competencia y mezcla de las tres clases de estímulos audiovisuales no bastan para explicar la respuesta conductual reportada frente a estímulos McGurk (Figura 4.6); empero, los estímulos incongruentes generaron activaciones corregidas diferenciadas en el MMR(ver 4.1). Con base en esto, proponemos que el percepto mixto [*da*] es resultado de la resolución del error multimodal con base en la experiencia multimodal (i.e., cantidad de competidores y fidelidad de cada entrada sensorial); el efecto McGurk no es el resultado de la mezcla de la información auditiva y visual.

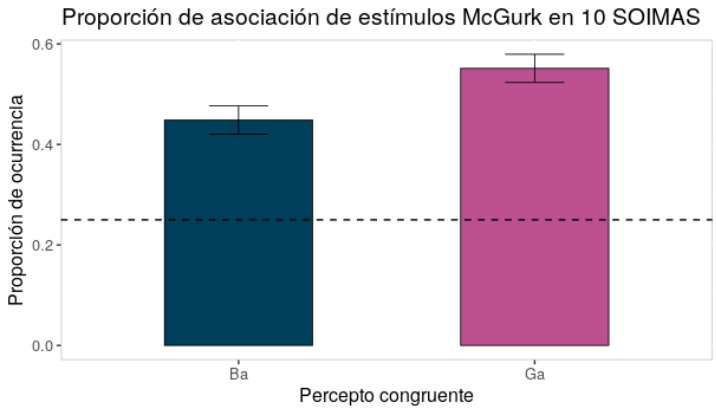

Figura 4.6: Proporción promedio de veces que unidades congruentes del MMR fuerron las unidades ganadoras de la activación corregida como consecuencia de estímulos McGurk. Podemos observar como la activación corregida ayuda a resolver la ambiguedad de los estímulos McGurk en favor de alguno de sus componentes modales ([*ga*] visual, [*ba*] auditivo). Los estímulos McGurk practicamente no fueron asociados a [*da*], esto es evidencia de el percepto [*da*] no emerge como consecuencia de la mezcla de información incongruente; sino, como una manera de resolver el error predictivo.

Capítulo 5
Discusión

En el presente trabajo utilizamos la arquitectura SOIMA (Escobar-Juárez et al., 2016) para comprobar que la mezcla de información incongruente en estímulos McGurk no ocaciona una mezcla de sus partes unimodales, sino un error multimodal que deberá de ser disminuido. Nuestra arquitectura posee la ventaja de estar basada procesos de dominio general como la autoorganización de sus elementos, así como el aprendizaje estadístico de las coocurrencias unimodales (Frost et al., 2015; Christiansen and Chater, 2015). Nuestra arquitectura nos permite argumentar a favor de que los procesos cognitivos de dominio general explican la percepción de información lingüística mediante la generación de predicciones multimodales.

Al igual que en estudios anteriores (Hisanaga et al., 2016; Deonarine et al., 2012), observamos cómo la integración multimodal facilita la comprensión de la información unimodal simultánea pues los estímulos incongruentes generan errores que son resueltos en favor de la información unimodal (ver Figura 4.5). Aúnado a lo anterior, observamos que la integración multimodal conlleva un mejoramiento sobre el procesamiento simplemente auditivo y visual, pues dentro de estos mapas las activaciones no son plenamente diferenciables (Figura 4.1);

Pese a lo anterior, los estímulos incongruentes generaron activaciones dentro del MMR diferenciables entre sí y también diferenciables con respecto de los estímulos congruentes (Figura 4.1); es decir, aunque los estímulos incongruentes fueron resueltos en favor de uno congruente, la activación no fue igual a ninguno de estos. Aúnado a esto, pudimos observar que distintas diadas incongruentes generaron activaciones diferenciables entre sí; es decir, las activaciones no son resultado de la mezcla de sus partes sino a la resulición del error en favor de la entrada sensorial más fidedigna (i.e.,los estímulos McGurk ([*ba*]+[*ga*]) no ocacionan la misma activación que BaGa pese a que ambos mezclan *ba* y *ga* debido a

que *ba* es más fidedigna cuando se presenta visualmente que cuando se presenta auditivamente, Figura 4.1)

Pese a que existió variabilidad entre los SOIMAs sobre qué entrada sensorial era menos ambigüa, los análisis de BMU en 4.2 comprueban que los estímulos McGurk fueron constantemente asociados a sus componentes unimodales una cantidad de veces similar. Esto es evidencia de que la integración multimodal disminuye la ambigüedad perceptual al reducir el error.

Pese a los resultados de BMU, los resultados de similitud de los patrones de activación corregida confirman que ningún estímulo incongruente generó patrones de activaciones similares a los congruentes (varoles cercanos a .2 en una escala logarítmica de 0 a 1, Figura 4.3). El análisis demostró, en cambio, que todos los estímulos incongruentes generaron patrones de activación corregida que compartían muy poca información con respecto de los congruentes. Más aún, ninguno de los estímulos incongruentes generó activaciones corregidas en interacción con los estímulos congruentes; pese a ello, la BMU resuelve la incongruencia multimodal en favor de ellos.

Nuestros resultados fortalecen la postura teórica de que nuestro cerebro genera predicciones sobre la información multimodal que espera recibir con base en el aprendizaje estadístico de ocurrencias sensoriales (Wallace and Stein, 2007; Stein and Rowland, 2011; Frost et al., 2015; Morse and Cangelosi, 2017). Durante la percepción, las predicciones multimodales son comparadas con la información multimodal y se disminuye el error suscitado de acuerdo a la ambigüedad de las respectivas entradas sensoriales.

Los estímulos McGurk generan un error mayor que los estímulos congruentes y que los individuos buscan reducir de manera eficiente. En estudios de fMRI se ha observado que estímulos McGurk generan mayor actividad dentro del STS_l debido al error que suscitan en dicha área, lo que indirectamente ofrece evidencia congruente con nuestros resultados. Nuestra arquitectura podría ser util para recrear estudios de neuroimagen de integración multimodal y del efecto McGurk pues el MMR simula el funcionamiento del STS_l(Beauchamp, 2016; Nath and Beauchamp, 2012; Calvert et al., 2001; Van Engen et al., 2019).

Un análisis explotario de la base de datos de (Feng et al., 2019) con más de 300 gemelos monocigotos y heterocigotos demostró que entre 5

condiciones experimentales, en promedio se otorgaban 39 ($sd = 5.05$) respuestas diferentes de [*ba*], [*ga*] ,[*da*]; también obervamos que la media de proporción de [*da*] como percepto mixto es de 0.36 ($sd = 0.28$); es decir, es evidencia empírica de que los estímulos McGurk pueden ser percibidos de manera diferente a [*da*] y corroboramos que la aparición de dicho percepto mixto en promedio ocurre con poca frecuencia.

Los 10 SOIMAs analizados fueron entrenados con distintos ejemplares de información multimodal congruente ([*ba*], [*ga*] y [*da*]). Nuestra intención fue experimentar si, tal como se ha argumentado, se percibe [*da*] frente a estímulos McGurk debido a que es el elemento multimodal de la mezcla de *ga* visual y *ba* auditivo; si dicha afirmación fuese cierto, entrenar a los SOIMAs con las clases congruentes *ba*, *ga* y *da* debería de bastar para que *da* emerga como percepto mixto. Aunado a esto, el analizar varios SOIMAs nos permite replicar la posible variabilidad entre sujetos.

Confirmamos nuestra hipótesis de que [*da*] no es el percepto que mejor reduce el eror entre las predicciones multimodales y la integración multimodal. Asímismo brindamos evidencia sobre que la mezcla de la información sensorial no basta para explicar la emergencia de perceptos mixtos. Con base en esto, predecimos que dependiendo de la cantidad de clases competidoras y de la variabilidad individual debido a la experiencia; el percepto mixto que emerge en estímulos McGurk, se asemejará más a las respuestas conductuales observadas en humanos (Van Engen et al., 2019; Magnotti et al., 2018b; Basu Mallick et al., 2015; Magnotti et al., 2018a).

Nuestro trabajo aporta evidencia sobre que el procesamiento multimodal emplea predicciones basadas en el aprendizaje de coocurrencias multimodales. En nuestro SOIMA los mapas que representan procesos de percepción unimodal son poco eficaces para diferenciar clases; sin embargo, el procesamiento multimodal diferenció totalmente la información congruente como resultado del entrenamiento multimodal (Wallace and Stein, 1997; Stein and Rowland, 2019). El efecto McGurk es un resultado del procesamiento multimodal de información incongruente, por lo tanto, la manera en la cual se resuelve la incongruencia entre los estímulos unimodales depende directamente del desarrollo. La experiencia de aprendizaje de coocurrencias unimodales como facilitadora de predicciones multimodales determina la manera en que sean percibidos los estímulos incongruentes.

Finalmente, el éxito de nuestro enfoque para suscitar perceptos mixtos con elementos unimodales presentes en el lenguaje, aporta evidencia sobre que el lenguaje se basa en procesos generales de integración multimodal; por lo tanto, nuestros resultados sugerirían que el desarrollo del lenguaje en los humanos se debe a una fortaleza en el procesamiento multimodal y no a la emergencia de procesos especificos y únicos atribuibles al lenguaje.

Referencias

Basu Mallick, D., F. Magnotti, J., and S. Beauchamp, M. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin Review*, 22(5):1299–1307.

Beauchamp, M. S. (2016). Audiovisual Speech Integration. In *Neurobiology of Language*, pages 515–526. Elsevier.

Belpaeme, T. and Morse, A. F. (2012). Word and Category Learning in a Continous Semantic Domain: Comparing cross-situational and interactive learning. *Advances in Complex Systems*, 15(03n04):1250031.

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., and Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLOS ONE*, 13(11):e0207160.

Brown, V. A. and Strand, J. F. (2019). "Paying" attention to audiovisual speech: Do incongruent stimuli incur greater costs? *Attention, Perception, Psychophysics*.

Calvert, G. A., Hansen, P. C., Iversen, S. D., and Brammer, M. J. (2001). Detection of Audio-Visual Integration Sites in Humans by Application of Electrophysiological Criteria to the BOLD Effect. *NeuroImage*, 14(2):427–438.

Campbell, R., MacSweeney, M., Surguladze, S., Calvert, G., McGuire, P., Suckling, J., Brammer, M. J., and David, A. S. (2001). Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research*, 12(2):233–243.

Christiansen, M. H. and Chater, N. (2015). The language faculty that wasn't: a usage-based account of natural language recursion. *Frontiers in Psychology*, 6.

Conway, C. M. and Christiansen, M. H. (2005). Modality-Constrained Statistical Learning of Tactile, Visual, and Auditory Sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1):24–39.

Deonarine, J. M., Dawber, E. J., and Munhall, K. G. (2012). Sumby and Pollack revisted: The influence of live presentation on audiovisual speech perception. *The Journal of the Acoustical Society of America*, 132(3):2080–2080.

Drijvers, L., Vaitonytė, J., and Özyürek, A. (2019). Degree of Language Experience Modulates Visual Attention to Visible Speech and Iconic Gestures During Clear and Degraded Speech Comprehension. *Cognitive Science*, 43(10).

Escobar-Juárez, E., Schillaci, G., Hermosillo-Valadez, J., and Lara-Guzmán, B. (2016). A Self-Organized Internal Models Architecture for Coding Sensory–Motor Schemes. *Frontiers in Robotics and AI*, 3.

Feng, G., Zhou, B., Zhou, W., Beauchamp, M. S., and Magnotti, J. F. (2019). A Laboratory Study of the McGurk Effect in 324 Monozygotic and Dizygotic Twins. *Frontiers in Neuroscience*, 13(October):1–8.

Frost, R., Armstrong, B. C., Siegelman, N., and Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, 19(3):117–125.

Gentilucci, M. and Cattaneo, L. (2005). Automatic audiovisual integration in speech perception. *Experimental Brain Research*, 167(1):66–75.

Gold, B., Morgan, N., and Ellis, D. (2011). *Speech and Audio Signal Processing*. John Wiley Sons, Inc., Hoboken, NJ, USA, second edi edition.

Gustafsson, L., Jantvik, T., and Paplinski, A. P. (2014). A Self-organized artificial neural network architecture that generates the McGurk effect. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 3974–3980. IEEE.

Hauser, M. D. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598):1569–1579.

Hebb, D. O. (1962). *The organization of behavior: a neuropsychological theory*. Science Editions.

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., and Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports*, 8(1).

Hisanaga, S., Sekiyama, K., Igasaki, T., and Murayama, N. (2016). Language/Culture Modulates Brain and Gaze Processes in Audiovisual Speech Perception. *Scientific Reports*, 6(1):35265.

Hochett, C. F. and Hockett, C. D. (1960). The Origin of Speech. *Scientific American*, 203(3):88–97.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengthsin Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15.

Holler, J. and Levinson, S. C. (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Houde, J. F. and Nagarajan, S. S. (2011). Speech Production as State Feedback Control. *Frontiers in Human Neuroscience*, 5.

Kazemi, V. and Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874. IEEE.

Keil, J., Muller, N., Ihssen, N., and Weisz, N. (2012). On the Variability of the McGurk Effect: Audiovisual Integration Depends on Prestimulus Brain States. *Cerebral Cortex*, 22(1):221–231.

Keough, M., Derrick, D., and Gick, B. (2019). Cross-Modal Effects in Speech Perception. *Annual Review of Linguistics*, 5(1):49–66.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Lara, B., Astorga, D., Mendoza-Bock, E., Pardo, M., Escobar, E., and Ciria, A. (2018). Embodied Cognitive Robotics and the learning of sensorimotor schemes. *Adaptive Behavior*, 26(5):225–238.

Lienhart, R. and Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In *Proceedings. International Conference on Image Processing*, volumen 1, pages I–900–I–903. IEEE.

Liu, X., Cheung, Y.-m., and Tang, Y. Y. (2016). Lip event detection using oriented histograms of regional optical flow and low rank affinity pursuit. *Computer Vision and Image Understanding*, 148:153–163.

Lüttke, C. S., Ekman, M., van Gerven, M. A. J., and de Lange, F. P. (2016). McGurk illusion recalibrates subsequent auditory perception. *Scientific Reports*, 6(1):32891.

Magnotti, J. F., Basu Mallick, D., and Beauchamp, M. S. (2018a). Reducing Playback Rate of Audiovisual Speech Leads to a Surprising Decrease in the McGurk Effect. *Multisensory Research*, 31(1-2):19–38.

Magnotti, J. F., Basu Mallick, D., Feng, G., Zhou, B., Zhou, W., and Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental Brain Research*, 233(9):2581–2586.

Magnotti, J. F., Smith, K. B., Salinas, M., Mays, J., Zhu, L. L., and Beauchamp, M. S. (2018b). A causal inference explanation for enhancement of multisensory integration by co-articulation. *Scientific Reports*, 8(1):18032.

Mcgurk, H. and Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Miller, L. M. (2005). Perceptual Fusion and Stimulus Coincidence in the Cross-Modal Integration of Speech. *Journal of Neuroscience*, 25(25):5884–5893.

Morse, A., Belpaeme, T., Cangelosi, A., and Floccia, C. (2011). Modeling U Shaped Performance Curves in Ongoing Development. *submitted to the Cognitive Science Conference*, Volume 33:3034–3039.

Morse, A. F. and Cangelosi, A. (2017). Why Are There Developmental Stages in Language Learning? A Developmental Robotics Model of Language Development. *Cognitive Science*, 41:32–51.

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (2015). Audio-visual speech scene analysis: Characterization of the dynamics of unbinding and rebinding the McGurk effect. *The Journal of the Acoustical Society of America*, 137(1):362–377.

Nath, A. R. and Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage*, 59(1):781–787.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273.

Olasagasti, I., Bouton, S., and Giraud, A.-L. (2015). Prediction across sensory modalities: A neurocomputational model of the McGurk effect. *Cortex*, 68:61–75.

Omata, K. and Mogi, K. (2008). Fusion and combination in audio-visual integration. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 464(2090):319–340.

Sato, M. and Shiller, D. M. (2018). Auditory prediction during speaking and listening. *Brain and Language*, 187:92–103.

Stein, B. E. and Rowland, B. A. (2011). *Organization and plasticity in multisensory integration. Early and late experience affects its governing principles*, volumen 191. Elsevier B.V., 1 edition.

Stein, B. E. and Rowland, B. A. (2019). *Neural development of multisensory integration*. Elsevier Inc.

Strand, J., Cooperman, A., Rowe, J., and Simenstad, A. (2014). Individual Differences in Susceptibility to the McGurk Effect: Links With Lipreading and Detecting Audiovisual Incongruity. *Journal of Speech, Language, and Hearing Research*, 57(6):2322–2331.

Tian, X. and Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, 6.

Tourville, J. A. and Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and Cognitive Processes*, 26(7):952–981.

Twomey, K. E., Morse, A. F., Cangelosi, A., and Horst, J. S. (2016). Competition Affects Word Learning in a Developmental Robotic System. In *Neurocomputational Models of Cognitive Development and Processing*, pages 131–144. WORLD SCIENTIFIC.

Ursino, M., Cuppini, C., and Magosso, E. (2014). Neurocomputational approaches to modelling multisensory integration in the brain: A review. *Neural Networks*, 60:141–165.

Van Engen, K. J., Dey, A., Sommers, M., and Peelle, J. (2019). Audiovisual Speech perception: Moving beyond McGurk.

Van Engen, K. J., Xie, Z., and Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, Psychophysics*, 79(2):396–403.

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, 45(3):598–607.

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volumen 1, pages I–511–I–518. IEEE Comput. Soc.

Wallace, M. T. and Stein, B. E. (1997). Development of multisensory neurons and multisensory integration in cat superior colliculus. *Journal of Neuroscience*, 17(7):2429–2444.

Wallace, M. T. and Stein, B. E. (2007). Early experience determines how the senses will interact. *Journal of Neurophysiology*, 97(1):921–926.

Zhang, J., Meng, Y., He, J., Xiang, Y., Wu, C., Wang, S., and Yuan, Z. (2019). McGurk Effect by Individuals with Autism Spectrum Disorder and Typically Developing Controls: A Systematic Review and Meta-analysis. *Journal of Autism and Developmental Disorders*, 49(1):34–43.

Zhen, B., Xihong, W., Zhimin, L., and HuishengChi (2000). On the Importance of Components of the MFCC in speech and speaker secognition. *ICSLP-2000*, 2:487–490.

Zhu, L. L. and Beauchamp, M. S. (2017). Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *The Journal of Neuroscience*, 37(10):2697–2708.